# Chapter 14

# Information Retrieval for Healthcare

**William R. Hersh**
*Department of Medical Informatics & Clinical Epidemiology (DMICE)*
*Oregon Health & Science University*
*Portland, OR*
hersh@ohsu.edu



## 14.1 Introduction

Although most work in Healthcare Data Analytics focuses on mining and analyzing data from patients, another vast trove of information for use in this process includes scientific data and literature. The techniques most commonly used to access this day include those from the field of *information retrieval* (IR), sometimes called *search*. IR is the field concerned with the acquisition, organization, and searching of knowledge-based information, which is usually defined as information derived and organized from observational or experimental research [60, 66]. Although IR in biomedicine traditionally concentrated on the retrieval of text from the biomedical literature, the purview of content covered has expanded to include newer types of media that include images,

video, chemical structures, gene and protein sequences, and a wide range of other digital media of relevance to biomedical education, research, and patient care. With the proliferation of IR systems and online content, even the notion of the library has changed substantially, with the new *digital library* emerging [90].

Figure 14.1 shows a basic overview of the IR process and forms the basis for most of this chapter. The overall goal of the IR process is to find *content* that meets a person's information needs. This begins with the posing of a *query* to the IR system. A *search engine* matches the query to content items through metadata. There are two intellectual processes of IR. *Indexing* is the process of assigning metadata to content items, while *retrieval* is the process of the user entering his or her query and retrieving content items.

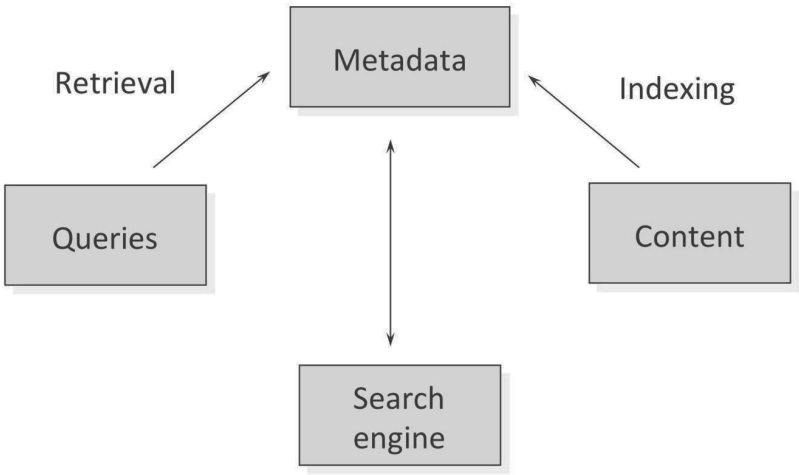


**FIGURE 14.1**: Basic overview of information retrieval (IR) process. (Copyright, William Hersh)

The use of IR systems has become essentially ubiquitous. It is estimated that among individuals who use the Internet in the United States, over 80 percent have used it to search for personal health information [50]. Virtually all physicians use the Internet [102]. Furthermore, access to systems has gone beyond the traditional personal computer and extended to new devices, such as smartphones and tablet devices.

Other evidence points to the importance of IR and biomedicine. One author now defines biology as an "information science" [76]. Another notes that pharmaceutical companies compete for informatics and library talent [28]. Clinicians can no longer keep up with the growth of the literature, as an average of 75 clinical trials and 11 systematic reviews are published each day [9]. *Search* is even part of the "meaningful use" program to incentivize adoption of the electronic health record, as text search over electronic notes is a requirement for obtaining incentive funding [94].

## 14.2 Knowledge-Based Information in Healthcare and Biomedicine

IR tends to focus on knowledge-based information, which is information based on scientific research and in distinction to patient-specific information that is generated in the care of the patient. Knowledge-based information is typically subdivided into two categories. Primary knowledge-based information (also called primary literature) is original research that appears in journals, books,

reports, and other sources. This type of information reports the initial discovery of health knowledge, usually with either original data or reanalysis of data (e.g., systematic reviews and meta-analyses). Secondary knowledge-based information consists of the writing that reviews, condenses, and/or synthesizes the primary literature. The most common examples of this type of literature are books, monographs, and review articles in journals and other publications. Secondary literature also includes opinion-based writing such as editorials and position or policy papers. It also encompasses clinical practice guidelines, narrative reviews, and health information on Web pages. In addition, it includes the plethora of pocket-sized manuals that were formerly a staple for practitioners in many professional fields. As will be seen later, secondary literature is the most common type of literature used by physicians. Secondary literature also includes the growing quality of patient/consumer-oriented health information that is increasingly available via the Web.

### 14.2.1 Information Needs and Seeking

It is important when designing IR systems to consider the needs of various users and the types of questions they bring to the system. Different users of knowledge-based information have differing needs based on the nature of what they need the information for and what resources are available. The information needs and information seeking of physicians have been most extensively studied. Gorman and Helfand [57] has defined four states of information need in the clinical context:

- Unrecognized need—clinician unaware of information need or knowledge deficit.
- Recognized need—clinician aware of need but may or may not pursue it.
- Pursued need—information seeking occurs but may or may not be successful.
- Satisfied need—information seeking successful.

Studies of physician information needs find that they are likely to pursue only a minority of unanswered questions. A variety of studies over several decades have demonstrated that physicians in practice have unmet information on the order of two questions for every three patients seen and only pursue answers for about 30 percent of these questions [25, 57, 39]. When answers to questions are actually pursued, these studies showed that the most frequent source for answers to questions was colleagues, followed by paper-based textbooks. Therefore, it is not surprising that barriers to satisfying information needs remain [40]. It is probably likely that physicians use electronic sources more now than were measured in these earlier studies, with the widespread use of the electronic health record (EHR) as well as the ubiquity of portable smartphones and tablets. One possible approach to lowering the barrier to knowledge-based information is to link it more directly with the context of the patient in the EHR [22].

The information needs of other users are less well-studied. As noted above, surveys find about 80 percent of all Internet users have searched for personal health information [50]. About 4.5 percent of all queries to Web search engines are health-related [43]. Analyses show that consumers tend to search on the following categories of topics [49]:

- Specific disease or medical problem—66%
- Certain medical treatment or procedure—56%
- Doctors or other health professionals—44%
- Hospitals or other medical facilities—36%
- Health insurance, private or government—33%
- Food safety or recalls—29%

- Environmental health hazards—22%
- Pregnancy and childbirth—19%
- Medical test results—16%

### 14.2.2 Changes in Publishing

Profound changes have taken place in the publishing of knowledge-based information in recent years. Virtually all scientific journals are published electronically now. In addition, there is great enthusiasm for electronic availability of journals, as evidenced by the growing number of titles to which libraries provide access. When available in electronic form, journal content is easier and more convenient to access. Furthermore, since most scientists have the desire for widespread dissemination of their work, they have incentive for their papers to be available electronically. Not only is there the increased convenience of redistributing reprints, but research has found that freely available on the Web have a higher likelihood of being cited by other papers than those that are not [12]. As citations are important to authors for academic promotion and grant funding, authors have an incentive to maximize the accessibility of their published work.

The technical challenges to electronic scholarly publication have been replaced by economic and political ones [69, 113]. Printing and mailing, tasks no longer needed in electronic publishing, comprised a significant part of the "added value" from publishers of journals. There is still however value added by publishers, such as hiring and managing editorial staff to produce the journals, and managing the peer review process. Even if publishing companies as they are known were to vanish, there would still be some cost to the production of journals. Thus, while the cost of producing journals electronically is likely to be less, it is not zero, and even if journal content is distributed "free," someone has to pay the production costs. The economic issue in electronic publishing, then, is who is going to pay for the production of journals [113]. This introduces some political issues as well. One of them centers around the concern that much research is publicly funded through grants from federal agencies such as the National Institutes of Health (NIH) and the National Science Foundation (NSF). In the current system, especially in the biomedical sciences (and to a lesser extent in other sciences), researchers turn over the copyright of their publications to journal publishers. The political concern is that the public funds the research and the universities carry it out, but individuals and libraries then must buy it back from the publishers to whom they willingly cede the copyright. This problem is exacerbated by the general decline in funding for libraries.

Some proposed models of "open access" scholarly publishing keep the archive of science freely available [98, 121, 129]. The basic principle of open access publishing is that authors and/or their institutions pay the cost of production of manuscripts up front after they are accepted through a peer review process. After the paper is published, it becomes freely available on the Web. Since most research is usually funded by grants, the cost of open access publishing should be included in grant budgets. The uptake of publishers adhering to the open access model has been modest, with the most prominent being *Biomed Central* (BMC,`www.biomedcentral.com`) and the *Public Library of Science* (PLoS, `www.plos.org`).

Another model that has emerged is *PubMed Central* (PMC, `pubmedcentral.gov`). PMC is a repository of life science research articles that provides free access while allowing publishers to maintain copyright and even optionally keep the papers housed on their own servers. A lag time of up to 6 months is allowed so that journals can reap the revenue that comes with initial publication. The National Institutes of Health (NIH, `www.nih.gov`) now requires all research funded by its grants to be submitted to PMC, either in the form published by publishers or as a PDF of the last manuscript prior to journal acceptance (`publicaccess.nih.gov`). Publishers have expressed concern that copyrights give journals more control over the integrity of the papers they publish [35]. An alternative approach advocated by non-commercial (usually professional society) publishers is the

*DC Principles for Free Access to Science* (www.dcprinciples.org), which advocates reinvestment of revenues in support of science, use of open archives such as PMC as allowed by business constraints, commitment to some free publication, more open access for low-income countries, and no charges for authors to publish.

## 14.3 Content of Knowledge-Based Information Resources

The previous sections of this chapter have described some of the issues and concerns surrounding the production and use of knowledge-based information in biomedicine. It is useful to classify the information to gain a better understanding of its structure and function. In this section, we classify content into bibliographic, full-text, annotated, and aggregated categories, although some content does not neatly fit within them.

### 14.3.1 Bibliographic Content

The first category consists of bibliographic content. It includes what was for decades the mainstay of IR systems: literature reference databases. Also called bibliographic databases, this content consists of citations or pointers to the medical literature (i.e., journal articles). The best-known and most widely used biomedical bibliographic database is MEDLINE, which contains bibliographic references to all of the biomedical articles, editorials, and letters to the editors in approximately 5,000 scientific journals. The journals are chosen for inclusion by an advisory committee of subject experts convened by NIH. At present, about 750,000 references are added to MEDLINE yearly. It now contains over 22 million references. A Web page devoted to MEDLINE size and searches statistics is at https://www.nlm.nih.gov/bsd/bsd_key.html.

The MEDLINE record may contain up to 49 fields. A user wanting just an overview on a topic may be interested in just a handful of these fields, such as the title, abstract, and indexing terms. But other fields contain specific information that may be of great importance to other audiences. For example, a genome researcher might be highly interested in the Supplementary Information (SI) field to link to genomic databases. A clinician may, however, derive benefit from some of the other fields. For example, the Publication Type (PT) field can help in the application of EBM, such as when one is searching for a practice guideline or a randomized controlled trial. MEDLINE is accessible by many means and available without charge via the PubMed system (http://pubmed.gov), produced by the National Center for Biotechnology Information (NCBI, www.ncbi.nlm.nih.gov) of the NLM, which provides access to other databases as well. A number of other information vendors, such as Ovid Technologies (www.ovid.com) and Aries Systems (www.ariessys.com), license the content of MEDLINE and other databases and provide value-added services that can be accessed for a fee by individuals and institutions.

MEDLINE is only one of many databases produced by the NLM. Other more specialized databases are also available, including textbooks, gene sequences, protein structures, and so forth. There are several non-NLM bibliographic databases that tend to be more focused on subjects or resource types. The major non-NLM database for the nursing field is the *Cumulative Index to Nursing and Allied Health Literature* (CINAHL, CINAHL Information Systems, http://www.ebscohost.com/cinahl/), which covers nursing and allied health literature, including physical therapy, occupational therapy, laboratory technology, health education, physician assistants, and medical records.

Another well-known bibliographic database is EMBASE (www.embase.com), which is sometimes referred to as the "European MEDLINE." It contains over 24 million records and covers many

of the same medical journals as MEDLINE but with a more international focus, including more non-English-language journals. These journals are often important for those carrying out meta-analyses and systematic reviews, which need access to all the studies done across the world.

A second, more modern type of bibliographic content is the Web catalog. There are increasing numbers of such catalogs, which consist of Web pages containing mainly links to other Web pages and sites. It should be noted that there is a blurry distinction between Web catalogs and aggregations (the fourth category). In general, the former contain only links to other pages and sites, while the latter include actual content that is highly integrated with other resources. Some well-known Web catalogs include:

- *HealthFinder* (www.healthfinder.gov)–consumer-oriented health information maintained by the Office of Disease Prevention and Health Promotion of the U.S. Department of Health and Human Services.
- *HON Select* (www.hon.ch/HONselect)–a European catalog of quality-filtered, clinician-oriented Web content from the HON foundation.
- *Translating Research into Practice* (TRIP, www.tripdatabase.com)–a database of content deemed to meet high standards of EBM.
- *Open Directory* (www.dmoz.org)–a general Web catalog that has significant health content.

An additional modern bibliographic resource is the *National Guidelines Clearinghouse* (NGC, www.guideline.gov). Produced by the Agency for Healthcare Research and Quality (AHRQ), it contains exhaustive information about clinical practice guidelines. Some of the guidelines produced are freely available, published electronically, and/or on paper. Others are proprietary, in which case a link is provided to a location at which the guideline can be ordered or purchased. The overall goal of the NGC is to make evidence-based clinical practice guidelines and related abstract, summary, and comparison materials widely available to healthcare and other professionals.

A final kind of bibliographic-like content consists of RSS feeds, which are short summaries of Web content: typically news, journal articles, blog postings, and other content. Users set up an RSS aggregation, which can be through a Web browser, email client, or standalone software, configured for the RSS feed desired, with an option to add a filter for specific content. There are two versions of RSS (1.0 and 2.0) but both provide:

- Title–name of item
- Link–URL to content
- Description–a brief description of the content

### 14.3.2 Full-Text Content

The second type of content is full-text content. A large component of this content consists of the online versions of books and periodicals. As already noted, most traditionally paper-based medical literature, from textbooks to journals, is now available electronically. The electronic versions may be enhanced by measures ranging from the provision of supplemental data in a journal article to linkages and multimedia content in a textbook. The final component of this category is the Web site. Admittedly, the diversity of information on Web sites is enormous, and sites may include every other type of content described in this chapter. However, in the context of this category, "Web site" refers to the vast number of static and dynamic Web pages at a discrete Web location.

Electronic publication of journals allows additional features not possible in the print world. Journal Web sites may provide supplementary data of results, images, and even raw data. A journal

Web site also allows more dialog about articles than could be published in a "Letters to the Editor" section of a print journal. Electronic publication also allows true bibliographic linkages, both to other full-text articles and to the MEDLINE record.

The Web also allows linkage directly from bibliographic databases to full text. PubMed maintains a field for the Web address of the full-text paper. This linkage is active when the PubMed record is displayed, but users may be met by a "paywall" if the article is not available for free. Many sites allow both access to subscribers or a pay-per-view facility. Many academic organizations now maintain large numbers of subscriptions to journals available to faculty, staff, and students. Other publishers, such as Ovid and MDConsult (www.mdconsult.com), provide access within their own password-protected interfaces to articles from journals that they have licensed for use in their systems.

The most common secondary literature source is traditional textbooks, which have essentially made a complete transition to publication in electronic form. A common approach with textbooks is bundling them, sometimes with linkages across the bundled texts. An early bundler of textbooks was Stat!-Ref (Teton Data Systems, www.statref.com) that, like many, began as a CD-ROM product and then moved to the Web. Stat!-Ref offers over 30 textbooks. Most other publishers have similarly aggregated their libraries of textbooks and other content. Another collection of textbooks is the NCBI Bookshelf, which contains many volumes on biomedical research topics (http://www.ncbi.nlm.nih.gov/books). One textbook that was formerly produced by NCBI but now is a standalone Web site is Online Mendelian Inheritance in Man (OMIM, http://omim.org), which is continually updated with new information about the genomic causes of human disease.

Electronic textbooks offer additional features beyond text from the print version. While many print textbooks do feature high-quality images, electronic versions offer the ability to have more pictures and illustrations. They also have the ability to provide sound and video. As with full-text journals, electronic textbooks can link to other resources, including journal references and the full articles. Many Web-based textbook sites also provide access to continuing education self-assessment questions and medical news. Finally, electronic textbooks let authors and publishers provide more frequent updates of the information than is allowed by the usual cycle of print editions, where new versions come out only every 2 to 5 years.

As noted above, Web sites are another form of full-text information. Probably the most effective provider of Web-based health information is the U.S. government. Not only do they produce bibliographic databases, but the NLM, AHRQ, the National Cancer Institute (NCI), Centers for Disease Control (CDC), and others have also been innovative in providing comprehensive full-text information for healthcare providers and consumers. One example is the popular CDC Travel site (http://www.cdc.gov/travel/). Some of these will be described later as aggregations, since they provide many different types of resources.

A large number of commercial biomedical and health Web sites have emerged in recent years. On the consumer side, they include more than just collections of text; they also include interaction with experts, online stores, and catalogs of links to other sites. Among the best known of these are Intelihealth (www.intelihealth.com) and NetWellness (www.netwellness.com). There are also Web sites, either from medical societies or companies, that provide information geared toward healthcare providers, typically overviews of diseases, their diagnosis, and treatment; medical news and other resources for providers are often offered as well.

Other sources of online health-related content include encyclopedias, the body of knowledge, and Weblogs or blogs. A well-known online encyclopedia with a great deal of health-related information is Wikipedia, which features a distributed authorship process whose content has been found to be reliable [56, 99] and frequently shows up near the top in health-related Web searches [86]. A growing number of organizations have a body of knowledge, such as the American Health Information Management Association (AHIMA, http://library.ahima.org/bok/). Blogs tend to carry a stream of consciousness but often high-quality information is posted within them.

### 14.3.3 Annotated Content

The third category consists of annotated content. These resources are usually not stored as free-standing Web pages but instead are often housed in database management systems. This content can be further subcategorized into discrete information types:

- Image databases—collections of images from radiology, pathology, and other areas.
- Genomics databases—information from gene sequencing, protein characterization, and other genomic research.
- Citation databases—bibliographic linkages of scientific literature.
- EBM databases—highly structured collections of clinical evidence.
- Other databases—miscellaneous other collections.

A great number of biomedical image databases are available on the Web. These include:

- Visible Human–http://www.nlm.nih.gov/research/visible/visible_human.html
- Lieberman's eRadiology–http://eradiology.bidmc.harvard.edu
- WebPath–http://library.med.utah.edu/WebPath/webpath.html
- Pathology Education Instructional Resource (PEIR)–www.peir.net
- DermIS–www.dermis.net
- VisualDX–www.visualdx.com

Many genomics databases are available on the Web. The first issue each year of the journal *Nucleic Acids Research* (NAR) catalogs and describes these databases, and is now available by open access means [55]. NAR also maintains an ongoing database of such databases, the Molecular Biology Database Collection (http://www.oxfordjournals.org/nar/database/a/). Among the most important of these databases are those available from NCBI [111]. All their databases are linked among themselves, along with PubMed and OMIM, and are searchable via the GQuery system (http://www.ncbi.nlm.nih.gov/gquery/).

Citation databases provide linkages to articles that cite others across the scientific literature. The earliest citation databases were the *Science Citation Index* (SCI, Thomspon-Reuters) and *Social Science Citation Index* (SSCI, Thomspon-Reuters), which are now part of the larger *Web of Science*. Two well-known bibliographic databases for biomedical and health topics that also have citation links include SCOPUS (www.scopus.com) and Google Scholar (http://scholar.google.com). These three were recently compared for their features and coverage [80]. A final citation database of note is CiteSeer (http://citeseerx.ist.psu.edu/), which focuses on computer and information science, including biomedical informatics. Evidence-based medicine (EBM) databases are devoted to providing annotated evidence-based information. Some examples include:

- *The Cochrane Database of Systematic Reviews*–one of the original collections of systematic reviews (www.cochrane.org).
- *Clinical Evidence*–an "evidence formulary" (www.clinicalevidence.com).
- *UpToDate*–content centered around clinical questions (www.uptodate.com).
- *InfoPOEMS*–"patient-oriented evidence that matters" (www.infopoems.com).

- *ACP Smart Medicine* (formerly Physicians' Information and Education Resource, PIER) "practice guidance statements" for which every test and treatment has associated ratings of the evidence to support them (`pier.acponline.org`).

There is a growing market for a related type of evidence-based content in the form of clinical decision support order sets, rules, and health/disease management templates. Publishers include EHR vendors whose systems employ this content as well as other vendors such as Zynx (`www.zynxhealth.com`) and Thomson Reuters Cortellis (`http://cortellis.thomsonreuters.com`).

There are a variety of other annotated content. The `ClinicalTrials.gov` database began as a database of clinical trials sponsored by NIH. In recent years it has expanded its scope to a register of clinical trials [30, 82] and to containing actual results of trials [131, 130]. Another important database for researchers is NIH RePORTER (`http://projectreporter.nih.gov/reporter.cfm`), which is a database of all research funded by NIH.

### 14.3.4 Aggregated Content

The final category consists of aggregations of content from the first three categories. The distinction between this category and some of the highly linked types of content described above is admittedly blurry, but aggregations typically have a wide variety of different types of information serving the diverse needs of users. Aggregated content has been developed for all types of users from consumers to clinicians to scientists.

Probably the largest aggregated consumer information resource is *MedlinePlus* (`http://medlineplus.gov`) from the NLM. MedlinePlus includes all of the types of content previously described, aggregated for easy access to a given topic. MedlinePlus contains health topics, drug information, medical dictionaries, directories, and other resources. Each topic contains links to health information from the NIH and other sources deemed credible by its selectors. There are also links to current health news (updated daily), a medical encyclopedia, drug references, and directories, along with a preformed PubMed search, related to the topic.

Aggregations of content have also been developed for clinicians. Most of the major publishers now aggregate all of their content in packages for clinicians. Another aggregated resource for clinicians is *Merck Medicus* (`www.merckmedicus.com`), developed by the well-known publisher and pharmaceutical house, is available for free to all licensed U.S. physicians, and includes a number of well-known resources, including some described above.

Another well-known group of aggregations of content for genomics researchers is the model organism databases. These databases bring together bibliographic databases, full text, and databases of sequences, structure, and function for organisms whose genomic data have been highly characterized. One of the oldest and most developed model organism databases is the Mouse Genome Informatics resource (`www.informatics.jax.org`).

## 14.4 Indexing

As described at the beginning of the chapter, indexing is the process of assigning metadata to content to facilitate its retrieval. Most modern commercial content is indexed in two ways:

1. Manual indexing–where human indexers, usually using a controlled terminology, assign indexing terms and attributes to documents, often following a specific protocol.

2. Automated indexing–where computers make the indexing assignments, usually limited to breaking out each word in the document (or part of the document) as an indexing term.

Manual indexing is done most commonly for bibliographic databases and annotated content. In this age of proliferating electronic content, such as online textbooks, practice guidelines, and multimedia collections, manual indexing has become either too expensive or outright unfeasible for the quantity and diversity of material now available. Thus, there are increasing numbers of databases that are indexed only by automated means. Before covering these types of indexing in detail, let us first discuss controlled terminologies.

### 14.4.1 Controlled Terminologies

A controlled terminology contains a set of terms that can be applied to a task, such as indexing. When the terminology defines the terms, it is usually called a vocabulary. When it contains variants or synonyms of terms, it is also called a thesaurus. Before discussing actual terminologies, it is useful to define some terms. A concept is an idea or object that occurs in the world, such as the condition under which human blood pressure is elevated. A term is the actual string of one or more words that represent a concept, such as *Hypertension* or *High Blood Pressure*. One of these string forms is the preferred or canonical form, such as Hypertension in the present example. When one or more terms can represent a concept, the different terms are called synonyms.

A controlled terminology usually contains a list of terms that are the canonical representations of the concepts. If it is a thesaurus, it contains relationships between terms, which typically fall into three categories:

- Hierarchical–terms that are broader or narrower. The hierarchical organization not only provides an overview of the structure of a thesaurus but also can be used to enhance searching (e.g., MeSH tree explosions that add terms from an entire portion of the hierarchy to augment a search).
- Synonym–terms that are synonyms, allowing the indexer or searcher to express a concept in different words.
- Related–terms that are not synonymous or hierarchical but are somehow otherwise related. These usually remind the searcher of different but related terms that may enhance a search.

The MeSH terminology is used to manually index most of the databases produced by the NLM [23]. The latest version contains over 26,000 subject headings (the word MeSH uses for the canonical representation of its concepts). It also contains over 170,000 synonyms to those terms, which in MeSH jargon are called entry terms. In addition, MeSH contains the three types of relationships described in the previous paragraph:

- Hierarchical–MeSH is organized hierarchically into 16 trees, such as Diseases, Organisms, and Chemicals and Drugs.
- Synonym–MeSH contains a vast number of entry terms, which are synonyms of the headings.
- Related–terms that may be useful for searchers to add to their searches when appropriate are suggested for many headings.

The MeSH terminology files, their associated data, and their supporting documentation are available on the NLM's MeSH Web site (`http://www.nlm.nih.gov/mesh/`). There is also a browser that facilitates exploration of the terminology (`http://www.nlm.nih.gov/mesh/MBrowser.html`). Figure 14.2 shows a slice through the MeSH hierarchy for certain cardiovascular diseases.

There are features of MeSH designed to assist indexers in making documents more retrievable.

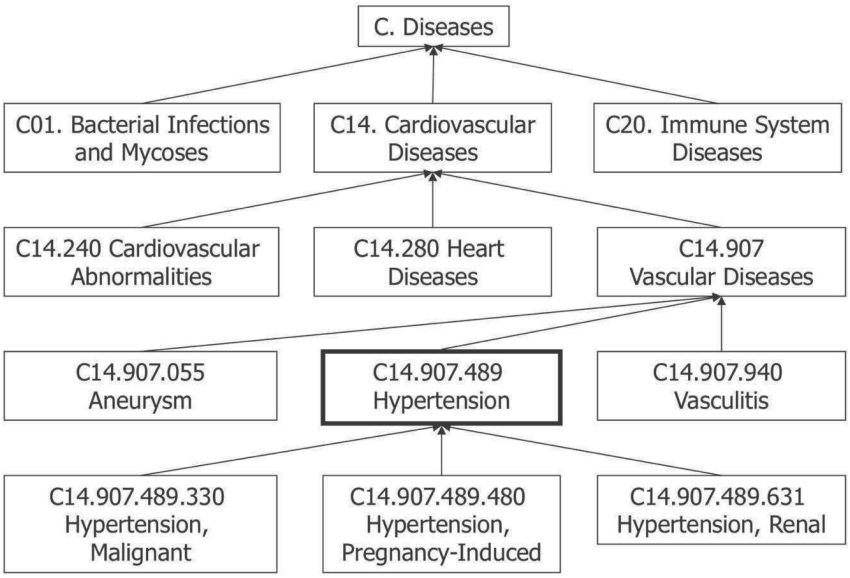


**FIGURE 14.2**: Portion of MeSH hierarchy for Cardiovascular Diseases. (Courtesy of NLM)

One of these is subheadings, which are qualifiers of subject headings that narrow the focus of a term. In Hypertension, for example, the focus of an article may be on the diagnosis, epidemiology, or treatment of the condition. Another feature of MeSH that helps retrieval is check tags. These are MeSH terms that represent certain facets of medical studies, such as age, gender, human or nonhuman, and type of grant support. Related to check tags are the geographical locations in the Z tree. Indexers must also include these, like check tags, since the location of a study (e.g., Oregon) must be indicated. Another feature gaining increasing importance for EBM and other purposes is the publication type, which describes the type of publication or the type of study. A searcher who wants a review of a topic may choose the publication type *Review* or *Review Literature*. Or, to find studies that provide the best evidence for a therapy, the publication type *Meta-Analysis, Randomized Controlled Trial,* or *Controlled Clinical Trial* would be used.

MeSH is not the only thesaurus used for indexing biomedical documents. A number of other thesauri are used to index non-NLM databases. CINAHL, for example, uses the CINAHL Subject Headings, which are based on MeSH but have additional domain-specific terms added. EMBASE has a terminology called EMTREE, which has many features similar to those of MeSH (http://www.embase.com/info/helpfiles/emtree-tool/emtree-thesaurus).

One problem with controlled terminologies, not limited to IR systems, is their proliferation. There is great need for linkage across these different terminologies. This was the primary motivation for the Unified Medical Language System (UMLS, http://www.nlm.nih.gov/research/umls/) Project, which was undertaken in the 1980s to address this problem [74]. There are three components of the UMLS Knowledge Sources: the Metathesaurus, the UMLS Semantic Network, and the Specialist Lexicon. The Metathesaurus component of the UMLS links parts or all of over 100 terminologies [13].

In the Metathesaurus, all terms that are conceptually the same are linked together as a concept. Each concept may have one or more terms, each of which represents an expression of the concept from a source terminology that is not just a simple lexical variant (i.e., differs only in word ending or order). Each term may consist of one or more strings, which represent all the lexical variants that are represented for that term in the source terminologies. One of each term's strings is designated as the preferred form, and the preferred string of the preferred term is known as the canonical form of

the concept. There are rules of precedence for determining the canonical form, the main one being that the MeSH heading is used if one of the source terminologies for the concept is MeSH.

Each Metathesaurus concept has a single concept unique identifier (CUI). Each term has one term unique identifier (LUI), all of which are linked to the one (or more) CUIs with which they are associated. Likewise, each string has one string unique identifier (SUI), which likewise are linked to the LUIs in which they occur. In addition, each string has an atomic unique identifier (AUI) that represents information from each instance of the string in each vocabulary. Figure 14.3 depicts the English-language concepts, terms, and strings for the Metathesaurus concept atrial fibrillation. (Each string may occur in more than one vocabulary, in which case each would be an atom.) The canonical form of the concept and one of its terms is atrial fibrillation. Within both terms are several strings, which vary in word order and case.

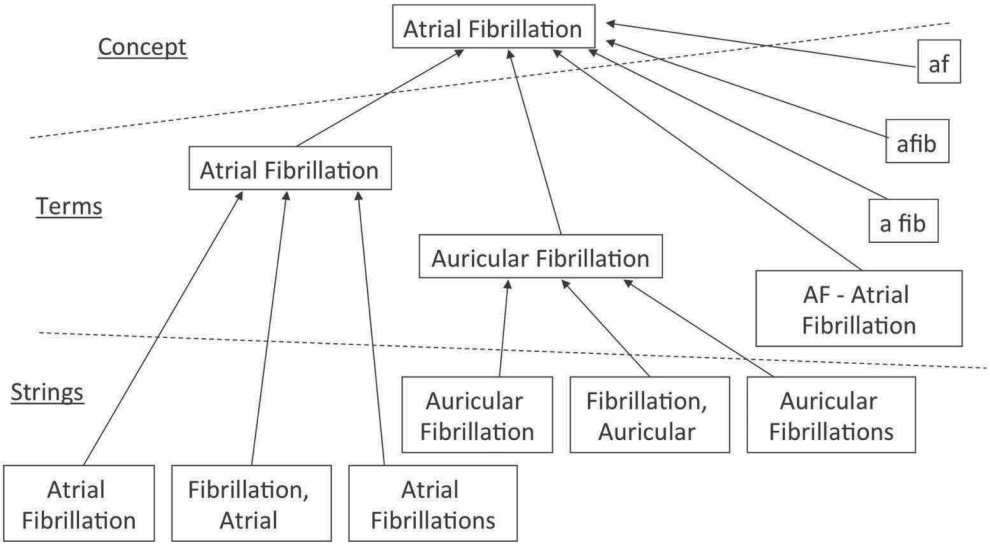


**FIGURE 14.3**: Unified Medical Language System Metathesaurus concept of atrial fibrillation. (Courtesy of NLM)

The Metathesaurus contains a wealth of additional information. In addition to the synonym relationships between concepts, terms, and strings described earlier, there are also nonsynonym relationships between concepts. There are a great many attributes for the concepts, terms, strings, and atoms, such as definitions, lexical types, and occurrence in various data sources. Also provided with the Metathesaurus is a word index that connects each word to all the strings it occurs in, along with its concept, term, string, and atomic identifiers.

### 14.4.2 Manual Indexing

Manual indexing is most commonly done for bibliographic and annotated content, although it is sometimes for other types of content as well. Manual indexing is usually done by means of a controlled terminology of terms and attributes. Most databases utilizing human indexing usually have a detailed protocol for assignment of indexing terms from the thesaurus. The MEDLINE database is no exception. The principles of MEDLINE indexing were laid out in the two-volume MEDLARS Indexing Manual [21]. Subsequent modifications have occurred with changes to MEDLINE, other databases, and MeSH over the years. The major concepts of the article, usually from two to five headings, are designed as main headings, and designated in the MEDLINE record by an asterisk. The indexer is also required to assign appropriate subheadings. Finally, the indexer must also as-

sign check tags, geographical locations, and publication types. Although MEDLINE indexing is still manual, indexers are aided by a variety of electronic tools for selecting and assigning MeSH terms.

Few full-text resources are manually indexed. One type of indexing that commonly takes place with full-text resources, especially in the print world, is that performed for the index at the back of the book. However, this information is rarely used in IR systems; instead, most online textbooks rely on automated indexing (see below). One exception to this is MDConsult, which uses back-of-book indexes to point to specific sections in its online books.

Manual indexing of Web content is challenging. With billions of pages of content, manual indexing of more than a fraction of it is not feasible. On the other hand, the lack of a coherent index makes searching much more difficult, especially when specific resource types are being sought. A simple form of manual indexing of the Web takes place in the development of the Web catalogs and aggregations as described earlier. These catalogs contain not only explicit indexing about subjects and other attributes, but also implicit indexing about the quality of a given resource by the decision of whether to include it in the catalog.

Two major approaches to manual indexing have emerged on the Web, which are often complementary. The first approach, that of applying metadata to Web pages and sites, is exemplified by the Dublin Core Metadata Initiative (DCMI, www.dublincore.org) [125]. The second approach, to build directories of content, was popularized initially by the Yahoo! search engine (www.yahoo.com). A more open approach to building directories was taken up by the Open Directory Project (www.dmoz.org), which carries on the structuring of the directory and entry of content by volunteers across the world.

The goal of the DCMI has been to develop a set of standard data elements that creators of Web resources can use to apply metadata to their content. The DCMI was recently approved as a standard by the National Information Standards Organization (NISO) with the designation Z39.85. It is also a standard with the International Organization for Standards (ISO), ISO Standard 15836:2009. The specification has 15 defined elements:

- DC.title—name given to the resource
- DC.creator—person or organization primarily responsible for creating the intellectual content of the resource
- DC.subject—topic of the resource
- DC.description—a textual description of the content of the resource
- DC.publisher—entity responsible for making the resource available in its present form
- DC.date—date associated with the creation or availability of the resource
- DC.contributor—person or organization not specified in a creator element who has made a significant intellectual contribution to the resource, but whose contribution is secondary to any person or organization specified in a creator element
- DC.type—category of the resource
- DC.format—data format of the resource, used to identify the software and possibly hardware that might be needed to display or operate the resource
- DC.identifier—string or number used to uniquely identify the resource
- DC.source—information about a second resource from which the present resource is derived
- DC.language—language of the intellectual content of the resource
- DC.relation—identifier of a second resource and its relationship to the present resource

- DC.coverage—spatial or temporal characteristics of the intellectual content of the resource
- DC.rights—rights management statement, an identifier that links to a rights management statement, or an identifier that links to a service providing information about rights management for the resource

There have been some medical adaptations of the DCMI. The most developed of these is the Catalogue et Index des Sites Mèdicaux Francophones (CISMeF, www.cismef.org) [27]. A catalog of French-language health resources on the Web, CISMeF has used DCMI to catalog over 40,000 Web pages, including information resources (e.g., practice guidelines, consensus development conferences), organizations (e.g., hospitals, medical schools, pharmaceutical companies), and databases. The Subject field uses the French translation of MeSH but also includes the English translation. For Type, a list of common Web resources has been enumerated.

While Dublin Core Metadata was originally envisioned to be included in Hypertext Markup Language (HTML) Web pages, it became apparent that many non-HTML resources exist on the Web and that there are reasons to store metadata external to Web pages. For example, authors of Web pages might not be the best people to index pages or other entities might wish to add value by their own indexing of content. A standard for cataloging metadata is the Resource Description Framework (RDF) [1]. A framework for describing and interchanging metadata, RDF is usually expressed in Extensible Markup Language (XML), a standard for data interchange on the Web. RDF also forms the basis of what some call the future of the Web as a repository not only of content but also of knowledge, which is also referred to as the Semantic Web [1]. Dublin Core Metadata (or any type of metadata) can be represented in RDF.

Manual indexing has a number of limitations, the most significant of which is inconsistency. Funk and Reid [54] evaluated indexing inconsistency in MEDLINE by identifying 760 articles that had been indexed twice by the NLM. The most consistent indexing occurred with check tags and central concept headings, which were only indexed with a consistency of 61 to 75 percent. The least consistent indexing occurred with subheadings, especially those assigned to noncentral concept headings, which had a consistency of less than 35 percent. A repeat of this study in more recent times found comparable results. Manual indexing also takes time. While it may be feasible with the large resources the NLM has to index MEDLINE, it is probably impossible with the growing amount of content on Web sites and in other full-text resources. Indeed, the NLM has recognized the challenge of continuing to have to index the growing body of biomedical literature and is investigating automated and semiautomated means of doing so [7].

### 14.4.3 Automated Indexing

In automated indexing, the indexing is done by a computer. Although the mechanical running of the automated indexing process lacks cognitive input, considerable intellectual effort may have gone into development of the system for doing it, so this form of indexing still qualifies as an intellectual process. In this section, we will focus on the automated indexing used in operational IR systems, namely the indexing of documents by the words they contain.

Some might not think of extracting all the words in a document as "indexing," but from the standpoint of an IR system, words are descriptors of documents, just like human-assigned indexing terms. Most retrieval systems actually use a hybrid of human and word indexing, in that the human-assigned indexing terms become part of the document, which can then be searched by using the whole controlled term or individual words within it. Most MEDLINE implementations have always allowed the combination of searching on human indexing terms and on words in the title and abstract of the reference. With the development of full-text resources in the 1980s and 1990s, systems that allowed only word indexing began to emerge. This trend increased with the advent of the Web.

Word indexing is typically done by defining all consecutive alphanumeric sequences between white space (which consists of spaces, punctuation, carriage returns, and other non-alphanumeric

characters) as words. Systems must take particular care to apply the same process to documents and the user's query, especially with characters such as hyphens and apostrophes. Many systems go beyond simple identification of words and attempt to assign weights to words that represent their importance in the document [107].

Many systems using word indexing employ processes to remove common words or conflate words to common forms. The former consists of filtering to remove *stop words*, which are common words that always occur with high frequency and are usually of little value in searching. The stop word list, also called a negative dictionary, varies in size from the seven words of the original MEDLARS stop list (and, an, by, from, of, the, with) to the list of 250 to 500 words more typically used. Examples of the latter are the 250-word list of van Rijsbergen, the 471-word list of Fox [48], and the PubMed stop list [3]. Conflation of words to common forms is done via *stemming*, the purpose of which is to ensure words with plurals and common suffixes (e.g., -ed, -ing, -er, -al) are always indexed by their stem form [51]. For example, the words cough, coughs, and coughing are all indexed via their stem cough. Both stop word remove and stemming reduce the size of indexing files and lead to more efficient query processing.

A commonly used approach for term weighting is $TF*IDF$ weighting, which combines the inverse document frequency (IDF) and term frequency (TF). The $IDF$ is the logarithm of the ratio of the total number of documents to the number of documents in which the term occurs. It is assigned once for each term in the database, and it correlates inversely with the frequency of the term in the entire database. The usual formula used is:

$$IDF(term) = log\frac{number\ of\ documents\ in\ database}{number\ of\ documents\ with\ term} + 1 \tag{14.1}$$

The $TF$ is a measure of the frequency with which a term occurs in a given document and is assigned to each term in each document, with the usual formula:

$$TF(term, document) = frequency\ of\ term\ in\ document \tag{14.2}$$

In $TF*IDF$ weighting, the two terms are combined to form the indexing weight, $WEIGHT$:

$$WEIGHT(term, document) = TF(term, document) * IDF(term) \tag{14.3}$$

Experiments from the Text Retrieval Conference (TREC, `trec.nist.gov`) (see section 14.6), led to the discovery of two other term-weighting approaches that have yielded consistently improved results. The first of these was based on a statistical model known as Poisson distributions and has been more commonly called *BM*25 weighting [105]. This weighting scheme is an improved document normalization approach, yielding up to 50% improvement in mean average precision (MAP) in various TREC collections [104]. One version of $TF$ for *BM*25 is:

$$BM25TF = \frac{(f_{td})(k_1+1)}{k_1(1-b) + k_1 b\dfrac{length\ of\ document}{average\ document\ length} + f_{td}} \tag{14.4}$$

$f_{td}$—frequency of terms in document

The variables $k_1$ and $b$ are parameters set to values based on characteristics of the collection. Typical values for $k_1$ are between 1 and 2 and for $b$ are between 0.6 and 0.75. A further simplification of this weighting often used is [104]:

$$BM25TF = \frac{(f_{td})}{0.5 + 1.5\dfrac{length\ of\ document}{average\ document\ length} + f_{td}} \tag{14.5}$$

$f_{td}$—-frequency of terms in document

Okapi weighting has its theoretical foundations in probabilistic IR, to be described shortly. As such, its $TF*IDF$ weighting uses a "probabilistic" variant of $IDF$:

$$BM25IDF = log\frac{t_d - number\ of\ documents\ with\ term + 0.5}{number\ of\ documents\ with\ term + 0.5} \tag{14.6}$$

$t_d$—total number of documents

The probabilistic model has also led to the newest theoretical approach to term weighting, known as language modeling, which will be described later in this section. Other techniques for term weighting have achieved varying amounts of success. One approach aimed to capture semantic equivalence of words in a document collection. Called latent semantic indexing (LSI), it uses a mathematically complex technique called singular-value decomposition (SVD) [31]. In LSI, an initial two-dimensional matrix of terms and documents is created, with the terms in one dimension and the documents in the other. The SVD process creates three intermediate matrices, the two most important being the mapping of the terms into an intermediate value, which can be thought to represent an intermediate measure of a term's semantics, and the mapping of this semantic value into the document. The number of intermediate values can be kept small, which allows the mapping of a large number of terms into a modest number of semantic classes or dimensions (i.e., several hundred). The result is that terms with similar semantic distributions (i.e., distributions that co-occur in similar document contexts) are mapped into the same dimension. Thus, even if a term does not co-occur with another, if it occurs in similar types of documents it will be likely to have similar semantics. While the optimal number of dimensions is not known, it has been shown for several of the small standard test collections that a few hundred is sufficient [31]. Some early evaluation studies showed small performance enhancements for LSI with small document collections [31, 73], but these benefits were not realized with larger collections such as TREC [36]. A better use for this technique may be with the automated discovery of synonymy [83].

Another approach to term weighting has been to employ probability theory. This approach is not necessarily at odds with the vector-space model, and in fact its weighting approaches can be incorporated into the vector-space model. The theory underlying probabilistic IR is a model to give more weight to terms likely to occur in relevant documents and unlikely to occur in nonrelevant documents. It is based on Bayes' theorem, a common probability measure that indicates likelihood of an event based on a prior situation and new data. Probabilistic IR is predominantly a relevance feedback technique, since some relevance information about the terms in documents is required. However, it did not show improvement over vector modification techniques in six older test collections [108]. In the TREC experiments, as noted earlier, some variants on the probabilistic approach were shown to perform better than vector-space relevance feedback with the addition of query expansion [15, 24, 81, 104, 124].

One modification to probabilistic IR was the inference model of Turtle and Croft [119], where documents were ranked based on how likely they are to infer belief they are relevant to the user's query. This method was also not necessarily incompatible with the vector-space model, and in some ways just provided a different perspective on the IR problem. One advantage of the inference model was the ability to combine many types of "evidence" that a document should be viewed by the user, such as queries with natural language and Boolean operators, as well as other attributes, such as citation of other documents. Combining some linguistic techniques, described later in this chapter, with slight modifications of TF*IDF weighting, passage retrieval, and query expansion, this approach performed consistently well in the TREC experiments [15].

A more recent application of probabilistic IR has been the use of language modeling [70]. This approach was adapted from other computer tasks, such as speech recognition and machine translation, where probabilistic principles are used to convert acoustic signals into words and words from one language to another, respectively. A key aspect of the language modeling approach is "smoothing" of the probabilities away from a purely deterministic approach of a term being present or

absent in a document in a binary fashion. Theoretically, the language modeling approach measures the probability of a query term given a relevant document.

Language modeling was introduced to the IR community by Ponte and Croft [101], who showed modest performance gains with TREC collections. A variety of enhancements were subsequently found to improve retrieval performance further [11]. Zhai and Lafferty [132] investigated smoothing models and derived a number of new conclusions about this approach to IR. Subsequent work processing text into topic signatures based on mapping to Unified Medical Language System (UMLS) Metathesaurus terms and using those instead of words found 10–20% performance gains with ad hoc retrieval data from the TREC Genomics Track [134].

Language models also allow the measurement of query "clarity," which is defined as a measure of the deviation between in the query and document language models from the general collection model [26]. Cronen-Townsend et al. found that query clarity was a good predictor of retrieval results from topics in the TREC ad hoc test collections, although application of this technique to real user queries from the TREC Interactive Track failed to uphold this association [118].

Another automated approach to pre-computing metadata about documents involves the use of link-based methods, which is best known through its use by the Google search engine (www.google.com). This approach gives weight to pages based on how often they are cited by other pages. The PageRank (PR) algorithm is mathematically complex, but can be viewed as giving more weight to a Web page based on the number of other pages that link to it [14]. Thus, the home page of the NLM or a major medical journal is likely to have a very high PR, whereas a more obscure page will have a lower PR. Google has also had to develop new computer architectures and algorithms to maintain pace with indexing the Web, leading to a new paradigm for such large-scale processing called MapReduce [29, 89].

In a simple description, *PR* can be viewed as giving more weight to a Web page based on the number of other pages that link to it. Thus, the home page of the NLM or JAMA is likely to have a very high *PR*, whereas a more obscure page will have a lower *PR*. The *PR* algorithm was developed by Brin and Page [14]. To calculate it for a given page *A*, it is assumed that there is a series of pages $T_1 \ldots T_n$ having links to *A*. There is another function C(A) that is the count of the number links going out of page *A*. There is also a "damping factor" *d* that is set between 0 and 1, by default at 0.85. Then *PR* is calculated for *A* as:

$$PR(A) = (1-d) + d\left(\frac{PR(T_1)}{C(T_1)} + \ldots + \frac{PR(T_n)}{C(T_n)}\right) \tag{14.7}$$

The algorithm begins by assigning every page a baseline value (such as the damping factor) and then iterates on a periodic basis. When implemented efficiently on a moderately-powered workstation, *PR* can be calculated for a large collection of Web pages.

It is often stated simplistically that *PR* is a form of measuring the in-degree, or the number of links, that point to a page. In reality, *PR* is more complex, giving added weight to pages that are pointed to by those that themselves have higher *PR*. Fortunato et al. [47] assessed how closely *PR* is approximated by simple in-degree, finding that the approximation was relatively accurate, allowing Web content creators to estimate their *PR* of their content by knowing the in-degree to their pages.

General-purpose search engines such as Google and Microsoft Bing (www.bing.com) use word-based approaches and variants of the PageRank algorithm for indexing. They amass the content in their search systems by "crawling" the Web, collecting and indexing every object they find on the Web. This includes not only HTML pages, but other files as well, including Microsoft Word, Portable Document Format (PDF), and images.

Word indexing has a number of limitations, including:

- Synonymy—different words may have the same meaning, such as high and elevated. This problem may extend to the level of phrases with no words in common, such as the synonyms *hypertension* and *high blood pressure*.

- Polysemy—the same word may have different meanings or senses. For example, the word *lead* can refer to an element or to a part of an electrocardiogram machine.
- Content—words in a document may not reflect its focus. For example, an article describing *hypertension* may make mention in passing to other concepts, such as *congestive heart failure* (CHF) that are not the focus of the article.
- Contex—words take on meaning based on other words around them. For example, the relatively common words *high, blood*, and *pressure*, take on added meaning when occurring together in the phrase *high blood pressure*.
- Morphology—words can have suffixes that do not change the underlying meaning, such as indicators of plurals, various participles, adjectival forms of nouns, and nominalized forms of adjectives.
- Granularity—queries and documents may describe concepts at different levels of a hierarchy. For example, a user might query for *antibiotics* in the treatment of a specific infection, but the documents might describe specific *antibiotics* themselves, such as *penicillin*.

A second purpose of indexing is to build structures so that computer programs can rapidly ascertain which documents use which indexing terms. Whether indexing is by terms in a thesaurus or words, IR systems are feasible only if they can rapidly process a user's query. A timely sequential search over an indexed text database is infeasible if not impossible for any large document collection. In IR, the usual approach involves the use of inverted files, where the terms are "inverted" to point to all the documents in which they occur. The algorithms for building and maintaining these structures have been used for decades [52]. An inverted file group for a sample document collection as it would be stored on a computer disk is shown in Figure 14.4. The first file is the *dictionary file*, which contains each indexing term along with a number representing how many documents contain

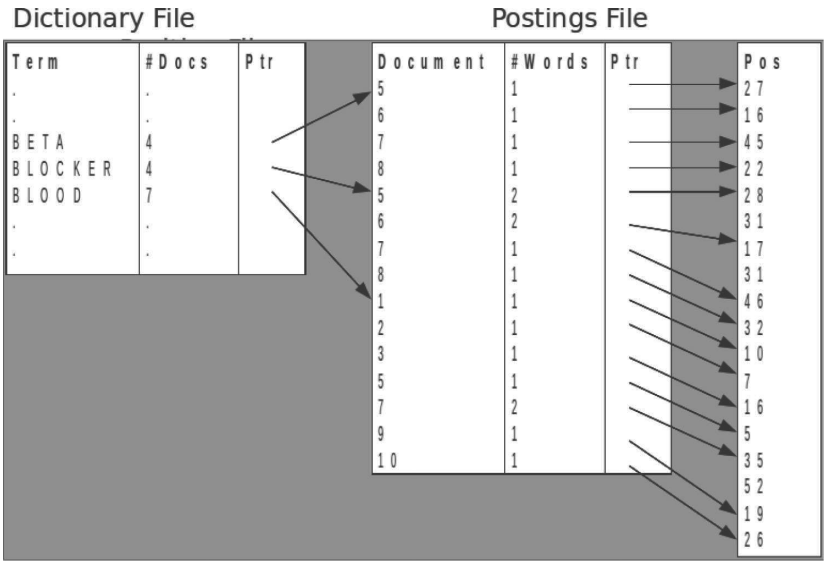


**FIGURE 14.4**: Inverted file structure used by information retrieval systems. Each term in the document collection occurs in the Dictionary File, which has a pointer to the Postings File, which has a pointer to each position of the word in the document in the Postings File. (Copyright, William Hersh)

the term and a pointer to the postings file. The *postings file* consists of a sequential list of all the documents that contain the indexing term. If it is desired to keep positional information for the indexing term (to allow proximity searching), then the postings file will also contain a pointer to the *position file*, which sequentially lists the positions of each indexing term in the document. The structure of the position file depends on what positional information is actually kept. The simplest position file contains just the word position within the document, while more complex files may contain the not only the word number, but also the sentence and paragraph number within the document.

The final component of inverted files is a mechanism for rapid lookup of terms in the dictionary file. This is typically done with a B-tree, which is a disk-based method for minimizing the number of disk accesses required to find a term in an index, resulting in fast lookup. The B-tree is very commonly used for keys in a DBMS. Another method for fast-term lookup is hashing [52].

Of course, with the need to process millions of queries each minute, just having an efficient file and look-up structure is not enough. Systems must be distributed across many servers in disparate geographic locations. Although the details of its approach are proprietary, Google has published some on how it maintains its subsecond response time to queries from around the globe [8, 29].



## 14.5 Retrieval

There are two broad approaches to retrieval. Exact-match searching allows the user precise control over the items retrieved. Partial-match searching, on the other hand, recognizes the inexact nature of both indexing and retrieval, and instead attempts to return the user content ranked by how close it comes to the user's query. After general explanations of these approaches, we will describe actual systems that access the different types of biomedical content.

### 14.5.1 Exact-Match Retrieval

In exact-match searching, the IR system gives the user all documents that exactly match the criteria specified in the search statement(s). Since the Boolean operators AND, OR, and NOT are usually required to create a manageable set of documents, this type of searching is often called Boolean searching. Furthermore, since the user typically builds sets of documents that are manipulated with the Boolean operators, this approach is also called set-based searching. Most of the early operational IR systems in the 1950s through the 1970s used the exact-match approach, even though Salton and McGill was developing the partial-match approach in research systems during that time [110]. In modern times, exact-match searching tends to be associated with retrieval from bibliographic and annotated databases, while the partial-match approach tends to be used with full-text searching.

Typically the first step in exact-match retrieval is to select terms to build sets. Other attributes, such as the author name, publication type, or gene identifier (in the secondary source identifier field of MEDLINE), may be selected to build sets as well. Once the search term(s) and attribute(s) have been selected, they are combined with the Boolean operators. The Boolean AND operator is typically used to narrow a retrieval set to contain only documents with two or more concepts. The Boolean OR operator is usually used when there is more than one way to express a concept. The Boolean NOT operator is often employed as a subtraction operator that must be applied to another set. Some systems more accurately call this the ANDNOT operator.

Some retrieval systems allow terms in searches to be expanded by using the wild-card character, which adds all words to the search that begin with the letters up until the wild-card character. This approach is also called truncation. Unfortunately, there is no standard approach to using wild-card characters, so syntax for them varies from system to system. PubMed, for example, allows a single

asterisk at the end of a word to signify a wild-card character. Thus, the query word can* will lead to the words cancer and Candid, among others, being added to the search.

### 14.5.2 Partial-Match Retrieval

Although partial-match searching was conceptualized very early, it did not see widespread use in IR systems until the advent of Web search engines in the 1990s. This is most likely because exact-match searching tends to be preferred by "power users" whereas partial-match searching is preferred by novice searchers. Whereas exact-match searching requires an understanding of Boolean operators and (often) the underlying structure of databases (e.g., the many fields in MEDLINE), partial-match searching allows a user to simply enter a few terms and start retrieving documents.

The development of partial-match searching is usually attributed to Salton and McGill [110], who pioneered the approach in the 1960s. Although partial-match searching does not exclude the use of non-term attributes of documents, and for that matter does not even exclude the use of Boolean operators (e.g., [109]), the most common use of this type of searching is with a query of a small number of words, also known as a natural language query. Because Salton's approach was based on vector mathematics, it is also referred to as the vector-space model of IR. In the partial-match approach, documents are typically ranked by their closeness of fit to the query. That is, documents containing more query terms will likely be ranked higher, since those with more query terms will in general be more likely to be relevant to the user. As a result this process is called relevance ranking. The entire approach has also been called lexical-statistical retrieval.

The most common approach to document ranking in partial-match searching is to give each a score based on the sum of the weights of terms common to the document and query. Terms in documents typically derive their weight from the TF*IDF calculation described above. Terms in queries are typically given a weight of one if the term is present and zero if it is absent. The following formula can then be used to calculate the document weight across all query terms:

$$\textit{Document weight} = \sum_{\textit{all query terms}} WT_q * WT_d \tag{14.8}$$

$WT_q$—Weight of terms in query
$WT_d$—Weight of terms in document

This may be thought of as a giant OR of all query terms, with sorting of the matching documents by weight. The usual approach is for the system to then perform the same stop word removal and stemming of the query that was done in the indexing process. (The equivalent stemming operations must be performed on documents and queries so that complementary word stems will match.)

One problem with TF*IDF weighting is that longer documents accumulate more weight in queries simply because they have more words. As such, some approaches "normalize" the weight of a document. The most common approach is cosine normalization:

$$\textit{Document weight} = \frac{\sum_{\textit{all query terms}} WT_q * WT_d}{\sqrt{\left(\sum_{\textit{all query terms}} WT_q^2\right) * \left(\sum_{\textit{all document terms}} WT_q^2\right)}} \tag{14.9}$$

$WT_q$—Weight of terms in query
$WT_d$—Weight of terms in document

A variety of other variations to the basic partial-matching retrieval approach have been developed. One important addition is *relevance feedback*, a feature allowed by the partial-match approach, permits new documents to be added to the output based on their similarity to those deemed relevant by the user. This approach also allows reweighting of relevant documents already retrieved to higher positions on the output list. The most common approach is the modified Rocchio equation

employed by Buckley et al. [17]. In this equation, each term in the query is reweighted by adding value for the term occurring in relevant documents and subtracting value for the term occurring in nonrelevant documents. There are three parameters, $\alpha$, $\beta$, *and* $\gamma$, which add relative value to the original weight, the added weight from relevant documents, and the subtracted weight from nonrelevant documents, respectively. In this approach, the query is usually expanded by adding a specified number of query terms (from none to several thousand) from relevant documents to the query. Each query term takes on a new value based on the following formula:

$$\begin{aligned} &\textit{New query weight} = \\ &\alpha * \textit{Original query weight} \\ &+ \beta * \frac{1}{\textit{Number of relevant documents}} * \sum_{\textit{All relevant documents}} \textit{Weight in document} \\ &- \gamma * \frac{1}{\textit{Number of Nonrelevant documents}} * \sum_{\textit{All non-relevant documents}} \textit{Weight in document} \end{aligned} \quad (14.10)$$

When the parameters, $\alpha$, $\beta$, *and* $\gamma$, are set to one, this formula simplifies to:

$$\begin{aligned} &\textit{New query weight} = \\ &\textit{Original query weight} \\ &+ \textit{Average term weight in relevant documents} \\ &- \textit{Average term weight in nonrelevant documents} \end{aligned} \quad (14.11)$$

A number of IR systems offer a variant of relevance feedback that finds similar documents to a specified one. PubMed allows the user to obtain "related articles" from any given one in an approach similar to relevance feedback but which uses a different algorithm [127]. A number of Web search engines allow users to similarly obtain related articles from a specified Web page.

One enduring successful retrieval technique has been *query expansion*, where the relevance feedback technique is used without relevance information. Instead, a certain number of top-ranking documents are assumed to be relevant and the relevance feedback approach is applied. Query expansion techniques have been shown to be among the most consistent methods to improve performance in TREC. In TREC-3, Buckley et al. [18] used the Rocchio formula with parameters 8, 8, and 0 (which perform less reweighting for expansion terms than in the relevance feedback experiments cited earlier) along with the addition of the top 500 terms and 10 phrases to achieve a 20% performance gain. Others in TREC have also shown benefit with this approach [42, 15, 18, 78, 104]. Additional work by Mitra et al. [95] has shown that use of manually created Boolean queries, passage-based proximity constraints (i.e., Boolean constraints must occur within 50–100 words), and term co-occurrences (i.e., documents are given more weight when query terms co-occur) improves MAP performance further still. The value of query expansion (and other approaches) has been verified by Buckley [16], who has constructed a table comparing different features of TREC systems with each year's ad hoc retrieval collection (p. 311).

Whether using exact-match or partial-match approaches, efficiency in merging sets of documents or sorting individual documents based on weighting is achieved through the use of inverted files described previously. The indexing terms can be rapidly found in the dictionary file, with document collections merged in Boolean operations and/or weighted in partial-matching operations in the postings file.

### 14.5.3 Retrieval Systems

There are many different retrieval interfaces, with some of the features reflecting the content or structure of the underlying database. As noted above, PubMed is the system at NLM that searches

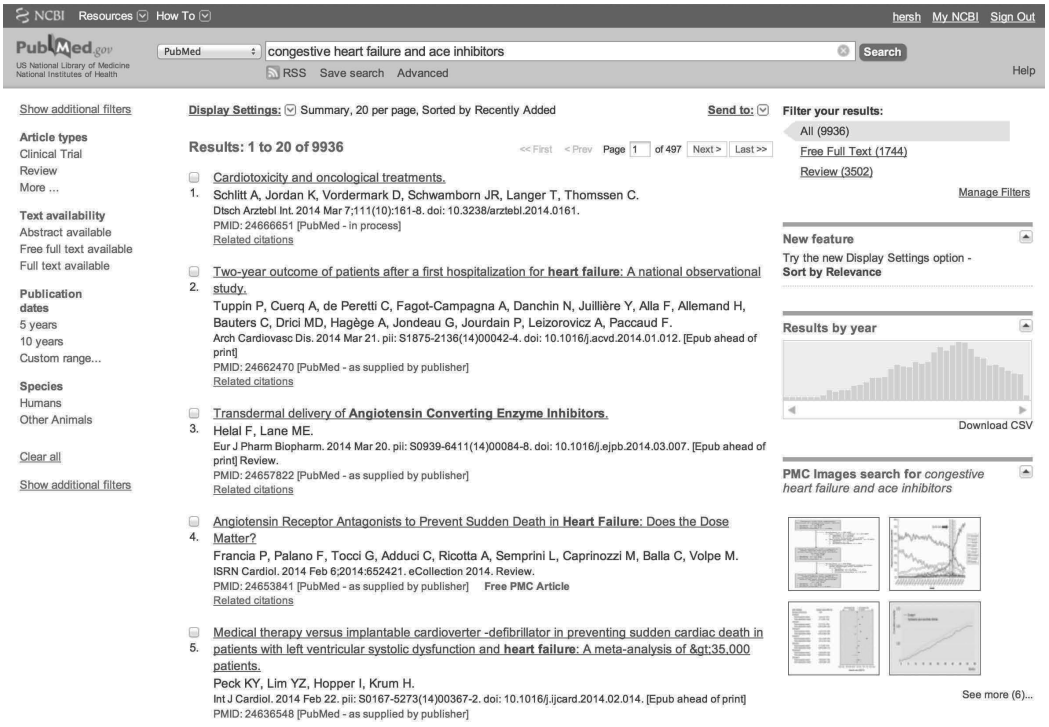


**FIGURE 14.5**: Screen shot of PubMed search. (Courtesy of NLM)

MEDLINE and other bibliographic databases. Although presenting the user with a simple text box, PubMed does a great deal of processing of the user's input to identify MeSH terms, author names, common phrases, and journal names (described in the online help system of PubMed). In this automatic term mapping, the system attempts to map user input, in succession, to MeSH terms, journals names, common phrases, and authors. Remaining text that PubMed cannot map is searched as text words (i.e., words that occur in any of the MEDLINE fields). Figure 14.5 shows the PubMed search results screen. The system allows a basic search and then provides access to a wealth of features around the results. The left-hand side of the screen allows setting of limits, such as to study type (e.g., randomized controlled trial), species (e.g., human or others), and age group (e.g., age >65 years). The right-hand side provides filters for free full-text article and reviews, as well as other features that include the details of the search. As in most bibliographic systems, users can search PubMed by building search sets and then combining them with Boolean operators to tailor the search. This is called the "advanced search" or "search builder" of PubMed, as shown in Figure 14.6. PubMed also has a specialized query interface for clinicians seeking the best clinical evidence (called Clinical Queries) as well as several "apps" that allow access via mobile devices (e.g., iOS or Android).

Another recent addition to PubMed is the ability to sort search results by relevance ranking rather than the long-standing default reverse-chronological ordering. Choosing this option leads to MEDLINE records being sorted based on a formula that includes IDF, TF, a measure for which field in which the word appears (more for title and abstract), and a measure of recency of publication [4].

As noted already, a great number of biomedical journals use the Highwire system for online access to their full text. The Highwire system provides a retrieval interface that searches over the complete online contents for a given journal. Users can search for authors, words limited to the title and abstract, words in the entire article, and within a date range. The interface also allows searching

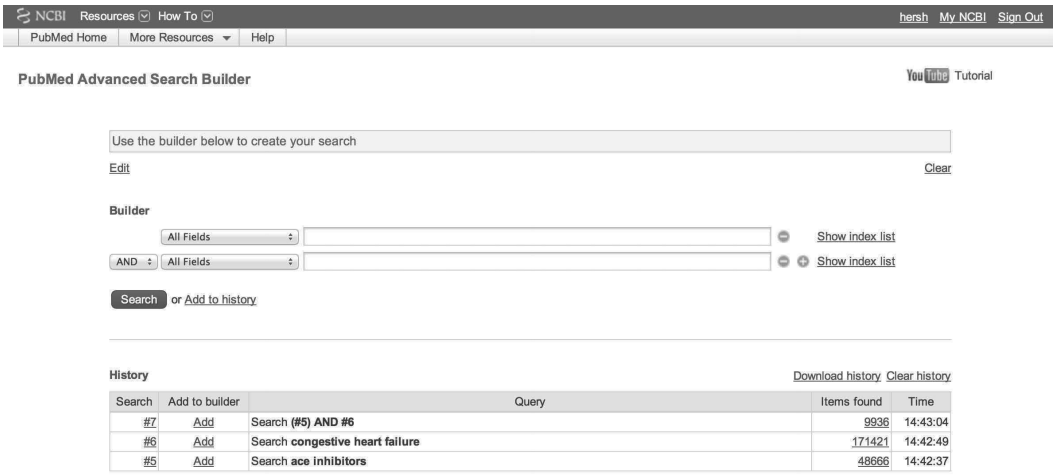


**FIGURE 14.6**: Screen shot of advanced search interface of PubMed. (Courtesy of NLM)



by citation by entering volume number and page as well as searching over the entire collection of journals that use Highwire. Users can browse through specific issues as well as collected resources.

Once an article has been found, a wealth of additional features is available. First, the article is presented both in HTML and PDF form, with the latter providing a more readable and printable version. Links are also provided to related articles from the journal as well as the PubMed reference and its related articles. Also linked are all articles in the journal that cited this one, and the site can be configured to set up a notification email when new articles cite the item selected. Finally, the Highwire software provides for "Rapid Responses," which are online letters to the editor. The online format allows a much larger number of responses than could be printed in the paper version of the journal. Other journal publishers use comparable approaches.

A growing number of search engines allow searching over many resources. The general search engines Google, Microsoft Bing, and others allow retrieval of any types of documents they index via their Web-crawling activities. Other search engines allow searching over aggregations of various sources, such as NLM's GQuery (`https://www.ncbi.nlm.nih.gov/gquery/gquery.fcgi`), which allows searching over all NLM databases and other resources in one simple interface.

## 14.6 Evaluation

There has been a great deal of research over the years devoted to the evaluation of IR systems. As with many areas of research, there is a controversy as to which approaches to evaluation best provide results that can assess searching in the systems they are using. Many frameworks have been developed to put the results in context. One of those frameworks organized evaluation around six questions that someone advocating the use of IR systems might ask [68]:

1. Was the system used?
2. For what was the system used?
3. Were the users satisfied?

4. How well did they use the system?
5. What factors were associated with successful or unsuccessful use of the system?
6. Did the system have an impact?

A simpler means for organizing the results of evaluation, however, groups approaches and studies into those which are system-oriented, i.e., the focus of the evaluation is on the IR system, and those which are user-oriented, i.e., the focus is on the user.

### 14.6.1 System-Oriented Evaluation

There are many ways to evaluate the performance of IR systems, the most widely used of which are the relevance-based measures of recall and precision. These measures quantify the number of relevant documents retrieved by the user from the database and in his or her search. Recall is the proportion of relevant documents retrieved from the database:

$$Recall = \frac{number\ of\ retrieved\ and\ relevant\ documents}{number\ of\ relevant\ documents\ in\ database} \tag{14.12}$$

In other words, recall answers the question, for a given search, what fraction of all the relevant documents have been obtained from the database?

One problem with Equation (14.5) is that the denominator implies that the total number of relevant documents for a query is known. For all but the smallest of databases, however, it is unlikely, perhaps even impossible, for one to succeed in identifying all relevant documents in a database. Thus, most studies use the measure of relative recall, where the denominator is redefined to represent the number of relevant documents identified by multiple searches on the query topic.

Precision is the proportion of relevant documents retrieved in the search:

$$Precision = \frac{number\ of\ retrieval\ and\ relevant\ documents}{number\ of\ documents\ retrieved} \tag{14.13}$$

This measure answers the question, for a search, what fraction of the retrieved documents are relevant?

One problem that arises when one is comparing systems that use ranking versus those that do not is that nonranking systems, typically using Boolean searching, tend to retrieve a fixed set of documents and as a result have fixed points of recall and precision. Systems with relevance ranking, on the other hand, have different values of recall and precision depending on the size of the retrieval set the system (or the user) has chosen to show. Often we seek to create an aggregate statistic that combines recall and precision. Probably the most common approach in evaluative studies is the mean average precision (MAP), where precision is measured at every point at which a relevant document is obtained, and the MAP measure is found by averaging these points for the whole query.

A good deal of evaluation in IR is done via challenge evaluations, where a common IR task is defined and a test collection of documents, topics, and relevance judgments are developed. The relevance judgments define which documents are relevant for each topic in the task, allowing different researchers to compare their systems with others on the same task and improve them. The longest running and best-known challenge evaluation in IR is the Text Retrieval Conference (TREC, `trec.nist.gov`), which is organized by the U.S. National Institute for Standards and Technology (NIST, `www.nist.gov`). Started in 1992, TREC has provided a testbed for evaluation and a forum for presentation of results. TREC is organized as an annual event at which the tasks are specified and queries and documents are provided to participants. Participating groups submit "runs" of their systems to NIST, which calculates the appropriate performance measure(s). TREC is organized into tracks geared to specific interests. A book summarizing the first decade of TREC grouped the tracks into general IR tasks [122]:

- Static text—ad hoc
- Streamed text—routing, filtering
- Human in the loop—interactive
- Beyond English (cross-lingual)—Spanish, Chinese, and others
- Beyond text—optical character recognition (OCR), speech, video
- Web searching—very large corpus
- Answers, not documents—question-answering
- Domain-specific—genomics, legal

While TREC has mostly focused on general-subject domains, there have been a couple of tracks that have focused on the biomedical domain. The first track to do so was the Genomics Track, which focused on the retrieval of articles as well as question-answering in this domain [64]. A second track to do focused on retrieval from medical records, with a task devoted to identifying patients who might be candidates for clinical studies based on criteria to be discerned from their medical records [123].

The TREC Genomics Track initially focused on improving MEDLINE retrieval. The ad hoc retrieval task modeled the situation of a user with an information need using an IR system to access the biomedical scientific literature. The document collection was based on a ten-year subset of MEDLINE. The rationale for using MEDLINE was that despite being in an era of readily available full-text journals (usually requiring a subscription), many users still entered the biomedical literature through searching MEDLINE. As such, there were still strong motivations to improve the effectiveness of searching MEDLINE.

The MEDLINE subset consisted of 10 years of completed citations from the database inclusive from 1994 to 2003. This provided a total of 4,591,008 records, which was about one-third of the full MEDLINE database. The data included all of the PubMed fields identified in the MEDLINE Baseline record. The size of the file uncompressed was about 9.5 gigabytes. In this subset, there were 1,209,243 (26.3%) records without abstracts.

Topics for the ad hoc retrieval task were based on information needs collected from real biologists. For both the 2004 and 2005 tracks, the primary measure of performance was MAP. Research groups were also required to classify their runs into one of three categories:

- Automatic—no manual intervention in building queries
- Manual—manual construction of queries but no further human interaction
- Interactive—completely interactive construction of queries and further interaction with system output

In the 2004 track, the best results were obtained by a combination of Okapi weighting (BM25 for term frequency but with standard inverse document frequency), Porter stemming, expansion of symbols by LocusLink and MeSH records, query expansion, and use of all three fields of the topic (title, need, and context) [53]. These achieved a MAP of 0.4075. When the language modeling technique of Dirichlet-Prior smoothing was added, an even higher MAP of 0.4264 was obtained. Another group achieved high-ranking results with a combination of approaches that included Okapi weighting, query expansion, and various forms of domain-specific query expansion (including expansion of lexical variants as well as acronym, gene, and protein name synonyms) [19]. Approaches that attempted to map to controlled vocabulary terms did not fare as well [6, 97, 112]. As always in

TREC, many groups tried a variety of approaches, beneficial or otherwise, but usually without comparing common baseline or running exhaustive experiments, making it difficult to discern exactly what techniques provided benefit.

Somewhat similar results were obtained in the 2005 track. As with 2004, the basic Okapi with good parameters gives good baseline performance for a number of groups. Manual synonym expansion of queries gave the highest MAP of 0.302 [72], although automated query expansion did not fare as well [2, 5]. Relevance feedback was found to be beneficial, but worked best without term expansion [133].

Follow-up research with the TREC Genomics Track ad hoc retrieval test collection has yielded a variety of findings. One study assessed word tokenization, stemming, and stop word removal, finding that varying strategies for the first resulted in substantial performance impact while changes in the latter two had minimal impact. Tokenization in genomics text can be challenging due to the use of a wide variety of symbols, including numbers, hyphens, super- and subscripts, and characters in non-English languages (e.g., Greek) [77].

Another TREC track focused on the biomedical domain was introduced in 2011 and run again in 2012, the TREC Medical Records Track [123]. The use case for the track TREC Medical Records Track was identifying patients from a collection of medical records who might be candidates for clinical studies. This is a real-world task for which automated retrieval systems could greatly aid in ability to carry out clinical research, quality measurement and improvement, or other "secondary uses" of clinical data [106]. The metric used to measure systems employed was inferred normalized distributed cumulative gain (infNDCG), which takes into account some other factors, such as incomplete judgment of all documents retrieval by all research groups.

The data for the track was a corpus of de-identified medical records developed by the University of Pittsburgh Medical Center. Records containing data, text, and ICD-9 codes are grouped by "visits" or patient encounters with the health system. (Due to the de-identification process, it was impossible to know whether one or more visits might emanate from the same patient.) There were 93,551 documents mapped into 17,264 visits.

A number of research groups used a variety of techniques, such as synonym and query expansion, machine learning algorithms, and matching against ICD-9 codes, but still had results that were not better than manually constructed queries employed by groups from NLM [32] or OHSU [10] (although the NLM system had a number of advanced features, such as document field searching [75]). Although the performance of systems in the track was "good" from an IR standpoint, they also showed that identification of patient cohorts would be a challenging task even for automated systems. Some of the automated features that had variable success included document section focusing, and term expansion, term normalization (mapping into controlled terms).

A number of approaches have been found to achieve modest improvement in results using data from this track. These include:

- Query expansion of normalized terms [103] and related terms [20, 79, 87]
- Detection of negation in records [88]
- Use of machine learning algorithms for ranking output [88, 135]

A failure analysis over the data from the 2011 track demonstrated why there are still many challenges that need to be overcome [37]. This analysis found a number of reasons why visits frequently retrieved were not relevant:

- Notes contain very similar term confused with topic
- Topic symptom/condition/procedure done in the past
- Most, but not all, criteria present

- All criteria present but not in the time/sequence specified by the topic description
- Topic terms mentioned as future possibility
- Topic terms not present–can't determine why record was captured
- Irrelevant reference in record to topic terms
- Topic terms denied or ruled out

The analysis also found reasons why visits rarely retrieval were actually relevant:

- Topic terms present in record but overlooked in search
- Visit notes used a synonym for topic terms
- Topic terms not named and must be derived
- Topic terms present in diagnosis list but not visit notes

Some researchers have criticized or noted the limitations of relevance-based measures. While no one denies that users want systems to retrieve relevant articles, it is not clear that the quantity of relevant documents retrieved is the complete measure of how well a system performs [115, 58]. Hersh [65] has noted that clinical users are unlikely to be concerned about these measures when they simply seek an answer to a clinical question and are able to do so no matter how many other relevant documents they miss (lowering recall) or how many nonrelevant ones they retrieve (lowering precision).

What alternatives to relevance-based measures can be used for determining performance of individual searches? Harter admits that if measures using a more situational view of relevance cannot be developed for assessing user interaction, then recall and precision may be the only alternatives. Some alternatives have focused on users being able to perform various information tasks with IR systems, such as finding answers to questions [38, 96, 61, 128, 63]. For several years, TREC featured an Interactive Track that had participants carry out user experiments with the same documents and queries [62]. Evaluations focusing on user-oriented evaluation of biomedical IR will be described in the next section.

### 14.6.2 User-Oriented Evaluation

A number of user-oriented evaluations have been performed over the years looking at users of biomedical information. Most of these studies have focused on clinicians.

One of the original studies measuring searching performance in clinical settings was performed by Haynes et al. [59]. This study also compared the capabilities of librarian and clinician searchers. In this study, 78 searches were randomly chosen for replication by both a clinician experienced in searching and a medical librarian. During this study, each original ("novice") user had been required to enter a brief statement of information need before entering the search program. This statement was given to the experienced clinician and librarian for searching on MEDLINE. All the retrievals for each search were given to a subject domain expert, blinded with respect to which searcher retrieved which reference. Recall and precision were calculated for each query and averaged. The results showed that the experienced clinicians and librarians achieved comparable recall in the range of 50%, although the librarians had better precision. The novice clinician searchers had lower recall and precision than either of the other groups. This study also assessed user satisfaction of the novice searchers, who despite their recall and precision results said that they were satisfied with their search outcomes. The investigators did not assess whether the novices obtained enough relevant articles to answer their questions, or whether they would have found additional value with the ones that were missed.

A follow-up study yielded some additional insights about the searchers [93]. As was noted, different searchers tended to use different strategies on a given topic. The different approaches replicated a finding known from other searching studies in the past, namely, the lack of overlap across searchers of overall retrieved citations as well as relevant ones. Thus, even though the novice searchers had lower recall, they did obtain a great many relevant citations not retrieved by the two expert searchers. Furthermore, fewer than 4 percent of all the relevant citations were retrieved by all three searchers. Despite the widely divergent search strategies and retrieval sets, overall recall and precision were quite similar among the three classes of users.

Recognizing the limitations of recall and precision for evaluating clinical users of IR systems, Hersh and co-workers [67] have carried out a number of studies assessing the ability of systems to help students and clinicians answer clinical questions. The rationale for these studies is that the usual goal of using an IR system is to find an answer to a question. While the user must obviously find relevant documents to answer that question, the quantity of such documents is less important than whether the question is successfully answered. In fact, recall and precision can be placed among the many factors that may be associated with ability to complete the task successfully.

The first study by this group using the task-oriented approach compared Boolean versus natural language searching in the textbook *Scientific American Medicine* [61]. Thirteen medical students were asked to answer 10 short-answer questions and rate their confidence in their answers. The students were then randomized to one or the other interface and asked to search on the five questions for which they had rated confidence the lowest. The study showed that both groups had low correct rates before searching (average 1.7 correct out of 10) but were mostly able to answer the questions with searching (average 4.0 out of 5). There was no difference in ability to answer questions with one interface or the other. Most answers were found on the first search to the textbook. For the questions that were incorrectly answered, the document with the correct answer was actually retrieved by the user two-thirds of the time and viewed more than half the time.

Another study compared Boolean and natural language searching of MEDLINE with two commercial products, CD Plus (now Ovid) and KF [63]. These systems represented the ends of the spectrum in terms of using Boolean searching on human-indexed thesaurus terms (Ovid) versus natural language searching on words in the title, abstract, and indexing terms (KF). Sixteen medical students were recruited and randomized to one of the two systems and given three yes/no clinical questions to answer. The students were able to use each system successfully, answering 37.5 percent correctly before searching and 85.4 percent correctly after searching. There were no significant differences between the systems in time taken, relevant articles retrieved, or user satisfaction. This study demonstrated that both types of systems can be used equally well with minimal training.

A more comprehensive study looked at MEDLINE searching by medical and nurse practitioner (NP) students to answer clinical questions. A total of 66 medical and NP students searched five questions each [67]. This study used a multiple-choice format for answering questions that also included a judgment about the evidence for the answer. Subjects were asked to choose from one of three answers:

- Yes, with adequate evidence.
- Insufficient evidence to answer question.
- No, with adequate evidence.

Both groups achieved a pre-searching correctness on questions about equal to chance (32.3 percent for medical students and 31.7 percent for NP students). However, medical students improved their correctness with searching (to 51.6 percent), whereas NP students hardly did at all (to 34.7 percent).

This study also attempted to measure what factors might influence searching. A multitude of factors, such as age, gender, computer experience, and time taken to search, were not associated with

successful answering of questions. Successful answering was, however, associated with answering the question correctly before searching, spatial visualization ability (measured by a validated instrument), searching experience, and EBM question type (prognosis questions easiest, harm questions most difficult). An analysis of recall and precision for each question searched demonstrated a complete lack of association with ability to answer these questions.

Two studies have extended this approach in various ways. Westbook et al. [126] assessed use of an online evidence system and found that physicians answered 37% of questions correctly before use of the system and 50% afterwards, while nurse specialists answered 18% of questions correctly and also 50% afterwards. Those who had correct answers before searching had higher confidence in their answers, but those not knowing the answer initially had no difference in confidence whether their answer turned out to be right or wrong. McKibbon and Fridsma [92] performed a comparable study of allowing physicians to seek answers to questions with resources they normally use employing the same questions as Hersh et al. [67]. This studies found no difference in answer correctness before or after using the search system. Clearly these study show a variety of effects with different IR systems, tasks, and users.

Pluye and Grad [100] performed a qualitative study assessing impact of IR systems on physician practice. The study identified 4 themes mentioned by physicians:

- Recall—of forgotten knowledge.
- Learning—new knowledge.
- Confirmation—of existing knowledge.
- Frustration—that system use not successful.

The researchers also noted two additional themes:

- Reassurance—that system is available.
- Practice improvement—of patient-physician relationship.

The bulk of more recent physician user studies have focused on ability to users to answer clinical questions. Hoogendam et al. compared UpToDate with PubMed for questions that arose in patient care among residents and attending physicians in internal medicine [71]. For 1305 questions, they found that both resources provided complete answers 53% of the time, but UpToDate was better at providing partial answers (83% full or partial answer for UpToDate compared to 63% full or partial answer for PubMed).

A similar study compared Google, Ovid, PubMed, and UpToDate for answering clinical questions among trainees and attending physicians in anaesthesiology and critical care medicine [117]. Users were allowed to select which tool to use for a first set of four questions to answer, while 1–3 weeks later they were randomized to only a single tool to answer another set of eight questions. For the first set of questions, users most commonly selected Google (45%), followed by UpToDate (26%), PubMed (25%), and Ovid (4.4%). The rate of answering questions correctly in the first set was highest for UpToDate (70%), followed by Google (60%), Ovid (50%), and PubMed (38%). The time taken to answer these questions was lowest for UpToDate (3.3 minutes), followed by Google (3.8 minutes), PubMed (4.4 minutes), and Ovid (4.6 minutes). In the second set of questions, the correct answer was most likely to be obtained by UpToDate (69%), followed by PubMed (62%), Google (57%), and Ovid (38%). Subjects randomized a new tool generally fared comparably, with the exception of those randomized from another tool to Ovid.

Another study compared searching UpToDate and PubMed Clinical Queries at the conclusion of a course for 44 medical residents in an information mastery course [41]. Subjects were randomized to one system for two questions and then the other system for another two questions. The correct answer was retrieved 76% of the time with UpToDate versus only 45% of the time with PubMed

Clinical Queries. Median time to answer the question was less for UpToDate (17 minutes) than PubMed Clinical Queries (29 minutes). User satisfaction was higher with UpToDate.

Fewer studies have been done assessing nonclinicians searching on health information. Lau et al. found that use of a consumer-oriented medical search engine that included PubMed, MedlinePLUS, and other resources by college undergraduates led to answers being correct at a higher rate after searching (82.0%) than before searching (61.2%) [85, 84]. Providing a feedback summary from prior searches boosted the success rate of using the system even higher, to 85.3%. Confidence in one's answer was not found to be highly associated with correctness of the answer, although confidence was likely to increase for those provided with feedback from other searchers on the same topic.

Despite the ubiquity of search systems, many users have skill-related problems when searching for information. van Duersen assessed a variety of computer-related and content-related skills from randomly selected subjects in the Netherlands [120]. Older age and lower educational level were associated with reduced skills, including use of search engines. While younger subjects were more likely to have better computer and searching skills than older subjects, they were more likely to use nonrelevant search results and unreliable sources in answering health-related questions. This latter phenomenon has also been seen outside the health domain among the "millennial" generation, sometimes referred to as "digital natives" [116].



## 14.7 Research Directions

The above evaluation research shows that there is still plenty of room for IR systems to improve their abilities. In addition, there will be new challenges that arise from growing amounts of information, new devices, and other new technologies.

There are also other areas related to IR where research is ongoing in the larger quest to help all involved in biomedicine and health–from patients to clinicians to researchers–better use information systems and technology to improve the application of knowledge to improve health. This has resulted in research taking place in a number of areas related to IR, which include:

- Information extraction and text mining–usually through the use of natural language processing (NLP) to extract facts and knowledge from text. These techniques are often employed to extract information from the EHR, with a wide variety of accuracy as shown in a recent systematic review [114]. Among the most successful uses of these techniques have been studies to identify diseases associated with genomic variations [33, 34].
- Summarization–Providing automated extracts or abstracts summarizing the content of longer documents [91, 46]
- Question-answering–Going beyond retrieval of documents to providing actual answers to questions, as exemplified by the IBM Corp. Watson system [44], which is being applied to medicine [45].

## 14.8 Conclusion

There has been considerable progress made in IR. Seeking online information is now done routinely not only by clinicians and researchers, but also by patients and consumers. There are still considerable challenges to make this activity more fruitful to users. They include:

- How do we lower the effort it takes for clinicians to get to the information they need rapidly in the busy clinical setting?
- How can researchers extract new knowledge from the vast quantity that is available to them?
- How can consumers and patients find high-quality information that is appropriate to their understanding of health and disease?
- Can the value added by the publishing process be protected and remunerated while making information more available?
- How can the indexing process become more accurate and efficient?
- Can retrieval interfaces be made simpler without giving up flexibility and power?

Although *search* has become a ubiquitous activity for many, there is still required research to answer these questions, move interaction to new devices, and discover how it will be implemented in the unforeseen advances in computing that will occur in the future.



## Bibliography

[1] R. Akerkar. *Foundations of the Semantic Web: XML, RDF & Ontology*. Alpha Science International, Ltd., 2009.

[2] R. K. Ando, M. Dredze, and T. Zhang. TREC 2005 Genomics Track experiments at IBM Watson. In *TREC*, 2014. http://trec.nist.gov/pubs/trec14/papers/ibm-tjwatson.geo.pdf

[3] Stopwords. In *PubMed Help*. National Library of Medicine, Bethesda, MD, 2007. http://www.ncbi.nlm.nih.gov/books/bv.fcgi?highlight=stopwords&rid=helppubmed.table.pubmedhelp.T43

[4] PubMed Help, 2014. http://www.ncbi.nlm.nih.gov/books/NBK3827/

[5] A. R. Aronson, D. Demner-Fushman, S. M. Humphrey, J. J. Lin, P. Ruch, M. E. Ruiz, L. H. Smith, L. K. Tanabe, W. J. Wilbur, and H. Liu. Fusion of knowledge-intensive and statistical approaches for retrieving and annotating textual genomics documents. In *TREC*, 2005. http://trec.nist.gov/pubs/trec14/papers/nlm-umd.geo.pdf

[6] A. R. Aronson, S. M. Humphrey, N. C. Ide, W. Kim, R. R. Loane, J. G. Mork, L. H. Smith, L. K. Tanabe, W. J. Wilbur, N. Xie, et al. Knowledge-intensive and statistical approaches to the retrieval and annotation of genomics MEDLINE citations. In *TREC*, 2004. http://trec.nist.gov/pubs/trec13/papers/nlm-umd-ul.geo.pdf

[7] A. R. Aronson, J. G. Mork, C. W. Gay, S. M. Humphrey, and W. J. Rogers. The NLM Indexing Initiative's Medical Text Indexer. *Medinfo*, 11(Pt 1):268–272, 2004.

[8] L. A. Barroso, J. Dean, and U. Holzle. Web search for a planet: The Google cluster architecture. *Micro, IEEE*, 23(2):22–28, 2003.

[9] H. Bastian, P. Glasziou, and I. Chalmers. Seventy-five trials and eleven systematic reviews a day: How will we ever keep up? *PLoS Medicine*, 7(9):e1000326, 2010.

[10] S. Bedrick, T. Edinger, A. Cohen, and W. Hersh. Identifying patients for clinical studies from electronic health records: TREC 2012 Medical Records Track at OHSU. In *The Twenty-First Text REtrieval Conference Proceedings (TREC 2012) [NIST Special Publication: SP 500-298]*. National Institute of Standards and Technology-NIST, 2012. http://trec.nist.gov/pubs/trec20/papers/OHSU.medical.update.pdf

[11] A. Berger and J. Lafferty. Information retrieval as statistical translation. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 222–229. ACM, 1999.

[12] B.-C. Björk and D. Solomon. Open access versus subscription journals: A comparison of scientific impact. *BMC Medicine*, 10(1):73, 2012. http://www.biomedcentral.com/1741-7015/10/73

[13] O. Bodenreider. The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl 1):D267–D270, 2004.

[14] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1):107–117, 1998.

[15] J. Broglio, J. Callan, W. Croft, and D. Nachbar. Document retrieval and routing using the Inquery system. In D. Harman, editor, *Overview of the Third Text REtrieval Conference (TREC-3)*, pages 29–38, Gaithersburg, MD, 1994. National Institute of Standards and Technology.

[16] C. Buckley. The Smart project at TREC. *Voorhees and Harman [2005]*, 2005.

[17] C. Buckley, G. Salton, and J. Allan. The effect of adding relevance information in a relevance feedback environment. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 292–300. Springer-Verlag New York, Inc., 1994.

[18] C. Buckley, G. Salton, J. Allan, and A. Singhal. Automatic query expansion using SMART: TREC 3. *NIST special publication*, pages 69–80, 1994.

[19] S. Buttcher, C. L. Clarke, and G. V. Cormack. Domain-specific synonym expansion and validation for biomedical information retrieval (Multitext experiments for TREC 2004). In *Proceedings of the 13th Text Retrieval Conference*, 2004.

[20] P. Callejas, A. Miguel, Y. Wang, and H. Fang. Exploiting domain thesaurus for medical record retrieval. In *The Twenty-First Text REtrieval Conference Proceedings (TREC 2012) [NIST Special Publication: SP 500-298]*. National Institute of Standards and Technology-NIST, 2012. http://trec.nist.gov/pubs/trec21/papers/udel_fang.medical.nb.pdf

[21] T. Charen. *MEDLARS Indexing Manual. Part 1: Bibliographic Principles and Descriptive Indexing*. National Library of Medicine, 1977.

[22] J. Cimino and G. Del Fiol. Infobuttons and point of care access to knowledge. *Clinical Decision Support—The Road Ahead*, pages 345–372, 2007.

[23] M. H. Coletti and H. L. Bleich. Medical Subject Headings used to search the biomedical literature. *Journal of the American Medical Informatics Association*, 8(4):317–323, 2001.

[24] W. Cooper, A. Chen, and F. Gey. Experiments in the probabilistic retrieval of documents. In D. Harman, editor, *Overview of the Third Text REtrieval Conference (TREC-3)*, pages 127–134, Gaithersburg, MD, 1994. National Institute of Standards and Technology.

[25] D. G. Covell, G. C. Uman, and P. R. Manning. Information needs in office practice: Are they being met? *Annals of Internal Medicine*, 103(4):596–599, 1985.

[26] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. Predicting query performance. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 299–306. ACM, 2002.

[27] S. Darmoni, J. Leroy, F. Baudic, M. Douyere, J. Piot, and B. Thirion. CISMEF: A structured health resource guide. *Methods of Information in Medicine*, 39(1):30–35, 2000.

[28] K. Davies. Search and deploy. Bio–IT World, October 16, 2006. http://www.bio-itworld.com/issues/2006/oct/biogen-idec/

[29] J. Dean and S. Ghemawat. MapReduce: Simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008.

[30] C. D. DeAngelis, J. M. Drazen, F. A. Frizelle, C. Haug, J. Hoey, R. Horton, S. Kotzin, C. Laine, A. Marusic, A. J. P. Overbeke, et al. Is this clinical trial fully registered?: A statement from the International Committee of Medical Journal Editors. *JAMA*, 293(23):2927–2929, 2005.

[31] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[32] D. Demner-Fushman, S. Abhyankar, A. Jimeno-Yepes, R. Loane, F. Lang, J. G. Mork, N. Ide, and A. R. Aronson. NLM at TREC 2012 medical records track. In *The Twenty-First Text REtrieval Conference Proceedings (TREC 2012) [NIST Special Publication: SP 500-298]*. National Institute of Standards and Technology-NIST, 2012. http://trec.nist.gov/pubs/trec21/papers/NLM.medical.final.pdf

[33] J. C. Denny. Mining electronic health records in the genomics era. *PLoS Computational Biology*, 8(12):e1002823, 2012. http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1002823

[34] J. C. Denny, L. Bastarache, M. D. Ritchie, R. J. Carroll, R. Zink, J. D. Mosley, J. R. Field, J. M. Pulley, A. H. Ramirez, E. Bowton, et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nature Biotechnology*, 31(12):1102–1111, 2013.

[35] J. M. Drazen and G. D. Curfman. Public access to biomedical research. *New England Journal of Medicine*, 351(13):1343–1343, 2004.

[36] S. T. Dumais et al. Latent semantic indexing (LSI): TREC-3 report. *Overview of the Third Text REtrieval Conference*, pages 219–230, 1994.

[37] T. Edinger, A. M. Cohen, S. Bedrick, K. Ambert, and W. Hersh. Barriers to retrieving patient information from electronic health record data: failure analysis from the TREC Medical Records Track. In *AMIA Annual Symposium Proceedings*, volume 2012, pages 180–188. American Medical Informatics Association, 2012.

[38] D. E. Egan, J. R. Remde, L. M. Gomez, T. K. Landauer, J. Eberhardt, and C. C. Lochbaum. Formative design evaluation of Superbook. *ACM Transactions on Information Systems (TOIS)*, 7(1):30–57, 1989.

[39] J. W. Ely, J. A. Osheroff, M. H. Ebell, G. R. Bergus, B. T. Levy, M. L. Chambliss, and E. R. Evans. Analysis of questions asked by family doctors regarding patient care. *BMJ*, 319(7206):358–361, 1999.

[40] J. W. Ely, J. A. Osheroff, M. H. Ebell, M. L. Chambliss, D. C. Vinson, J. J. Stevermer, and E. A. Pifer. Obstacles to answering doctors' questions about patient care with evidence: Qualitative study. *BMJ*, 324(7339):710, 2002.

[41] L. S. Ensan, M. Faghankhani, A. Javanbakht, S.-F. Ahmadi, and H. R. Baradaran. To compare PubMed clinical queries and UpToDate in teaching information mastery to clinical residents: A crossover randomized controlled trial. *PLoS ONE*, 6(8):e23487, 2011.

[42] D. Evans and R. Lefferts. Design and evaluation of the CLARIT TREC-2 system. *NIST special publication*, pages 137–150, 1993.

[43] G. Eysenbach and C. Köhler. Health-related searches on the internet. *JAMA*, 291(24):2946–2946, 2004.

[44] D. Ferrucci, E. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A. A. Kalyanpur, A. Lally, J. W. Murdock, E. Nyberg, J. Prager, et al. Building Watson: An overview of the DeepQA project. *AI Magazine*, 31(3):59–79, 2010.

[45] D. Ferrucci, A. Levas, S. Bagchi, D. Gondek, and E. T. Mueller. Watson: Beyond Jeopardy! *Artificial Intelligence*, 199–200:93–105, 2013.

[46] M. Fiszman, T. C. Rindflesch, and H. Kilicoglu. Summarization of an online medical encyclopedia. *Medinfo*, 11(Pt 1):506–510, 2004.

[47] S. Fortunato, M. Boguna, A. Flammini, and F. Menczer. How to make the top ten: Approximating PageRank from in-degree, 2005. http://arxiv.org/pdf/cs.IR/0511016

[48] C. Fox. Lexical analysis and stoplists. In *Information Retrieval*, pages 102–130. Prentice-Hall, Inc., 1992.

[49] S. Fox. Health topics. *Pew Internet & American Life Project*, 2011. http://www.pewinternet.org/Reports/2011/HealthTopics.aspx

[50] S. Fox and M. Duggan. Health online. *Health*, 2013. http://www.pewinternet.org/Reports/2013/Health-online.aspx

[51] W. Frakes. Stemming algorithms. In *Information Retrieval*, pages 131–160. Prentice-Hall, Inc., 1992.

[52] W. B. Frakes and R. Baeza-Yates. *Information Retrieval: Data Structures and Algorithms,* Prentice Hall, 1992.

[53] S. Fujita. Revisiting again document length hypotheses—TREC 2004 Genomics Track experiments at Patolis. In *TREC*, 2004. Available at http://trec.nist.gov/pubs/trec13/t13_proceedings.html

[54] M. Funk and C. Reid. Indexing consistency in MEDLINE. *Bulletin of the Medical Library Association*, 71(2):176–183, 1983.

[55] M. Y. Galperin and G. R. Cochrane. The 2011 nucleic acids research database issue and the online Molecular Biology Database Collection. *Nucleic Acids Research*, 39(suppl_1):D1–D6, 2011.

[56] J. Giles. Internet encyclopaedias go head to head. *Nature*, 438(7070):900–901, 2005.

[57] P. N. Gorman and M. Helfand. Information seeking in primary care how physicians choose which clinical questions to pursue and which to leave unanswered. *Medical Decision Making*, 15(2):113–119, 1995.

[58] S. P. Harter. Psychological relevance and information science. *Journal of the American Society for Information Science*, 43(9):602–615, 1992.

[59] R. B. Haynes, K. A. McKibbon, C. J. Walker, N. Ryan, D. Fitzgerald, and M. F. Ramsden. Online access to MEDLINE in clinical settings: A study of use and usefulness. *Annals of Internal Medicine*, 112(1):78–84, 1990.

[60] W. Hersh. *Information Retrieval: A Health and Biomedical Perspective.* Springer, 2009.

[61] W. Hersh and D. Hickam. An evaluation of interactive Boolean and natural language searching with an on-line medical textbook. *Journal of the American Society for Information Science*, 46:478–489, 1995.

[62] W. Hersh and P. Over. Interactivity at the Text REtrieval Conference (TREC). *Information Processing & Management*, 37(3):365–366, 2001.

[63] W. Hersh, J. Pentecost, and D. Hickam. A task-oriented approach to information retrieval evaluation. *Journal of the American Society for Information Science*, 47(1):50–56, 1996.

[64] W. Hersh and E. Voorhees. TREC Genomics special issue overview. *Information Retrieval*, 12(1):1–15, 2009.

[65] W. R. Hersh. Relevance and retrieval evaluation: perspectives from medicine. *Journal of the American Society for Information Science*, 45(3):201–206, 1994.

[66] W. R. Hersh. Information retrieval and digital libraries. In *Biomedical Informatics: Computer Applications in Healthcare and Biomedicine*, pages 613–641. Springer, 2014.

[67] W. R. Hersh, M. K. Crabtree, D. H. Hickam, L. Sacherek, C. P. Friedman, P. Tidmarsh, C. Mosbaek, and D. Kraemer. Factors associated with success in searching MEDLINE and applying evidence to answer clinical questions. *Journal of the American Medical Informatics Association*, 9(3):283–293, 2002.

[68] W. R. Hersh and D. H. Hickam. How well do physicians use electronic information retrieval systems? A framework for investigation and systematic review. *JAMA*, 280(15):1347–1352, 1998.

[69] W. R. Hersh and T. C. Rindfleisch. Electronic publishing of scholarly communication in the biomedical sciences. *Journal of the American Medical Informatics Association*, 7(3):324–325, 2000.

[70] D. Hiemstra and W. Kraaij. A Language-Modeling Approach to TREC. In *TREC: Experiment and Evaluation in Information Retrieval.* E. Voorhees and D. Harman (editors). MIT Press, Cambridge, MA. 2005.

[71] A. Hoogendam, A. F. Stalenhoef, P. F. de Vries Robbé, and A. J. P. Overbeke. Answers to questions posed during daily patient care are more likely to be answered by UpToDate than PubMed. *Journal of Medical Internet Research*, 10(4):e29–e29, 2008.

[72] X. Huang, M. Zhong, and L. Si York University at TREC 2005: Genomics Track. In *The Fourteenth Text REtrieval Conference Proceedings (TREC 2005)*. National Institute of Standards & Technology, 2005. http://trec.nist.gov/pubs/trec14/papers/yorku-huang2.geo.pdf

[73] D. Hull. Improving text retrieval for the routing problem using latent semantic indexing. In *SIGIR'94*, pages 282–291. Springer, 1994.

[74] B. L. Humphreys, D. A. Lindberg, H. M. Schoolman, and G. O. Barnett. The Unified Medical Language System: An informatics research collaboration. *Journal of the American Medical Informatics Association*, 5(1):1–11, 1998.

[75] N. C. Ide, R. F. Loane, and D. Demner-Fushman. Essie: A concept-based search engine for structured biomedical text. *Journal of the American Medical Informatics Association*, 14(3):253–263, 2007.

[76] T. R. Insel, N. D. Volkow, T.-K. Li, J. F. Battey Jr., and S. C. Landis. Neuroscience networks. *PLoS Biology*, 1(1):e17, 2003.

[77] J. Jiang and C. Zhai. An empirical study of tokenization strategies for biomedical information retrieval. *Information Retrieval*, 10(4-5):341–363, 2007.

[78] D. Knaus, E. Mittendorf, and P. Schäuble. Improving a basic retrieval method by links and passage level evidence. *TREC 3*, pages 241–246, 1994.

[79] B. Koopman, G. Zuccon, A. Nguyen, D. Vickers, L. Butt, and P. D. Bruza. Exploiting SNOMED CT concepts and relationships for clinical information retrieval: Australian e-health Research Centre and Queensland University of Technology at the TREC 2012 Medical Track. In *The Twenty-First Text REtrieval Conference Proceedings (TREC 2012)[NIST Special Publication: SP 500-298]*. National Institute of Standards and Technology-NIST, 2012. http://trec.nist.gov/pubs/trec21/papers/AEHRC.medical.nb.pdf

[80] A. V. Kulkarni, B. Aziz, I. Shams, and J. W. Busse. Comparisons of citations in Web of Science, Scopus, and Google Scholar for articles published in general medical journals. *JAMA*, 302(10):1092–1096, 2009.

[81] K. L. Kwok, L. Grunfeld, D. D. Lewis TREC-3 ad-hoc, routing retrieval, and thresholding experiments using PIRCS. *TREC 3*, pages 247–255, 1994.

[82] C. Laine, R. Horton, C. D. DeAngelis, J. M. Drazen, F. A. Frizelle, F. Godlee, C. Haug, P. C. Hébert, S. Kotzin, A. Marusic, et al. Clinical trial registration: Looking back and moving ahead. *JAMA*, 298(1):93–94, 2007.

[83] T. K. Landauer and S. T. Dumais. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240, 1997.

[84] A. Lau, T. Kwok, and E. Coiera. How online crowds influence the way individual consumers answer health questions–an online prospective study. *Applied Clinical Informatics*, 2:177–189, 2011.

[85] A. Y. Lau and E. W. Coiera. Impact of web searching and social feedback on consumer decision making: A prospective online experiment. *Journal of Medical Internet Research*, 10(1):e2–e2, 2008.

[86] M. R. Laurent and T. J. Vickers. Seeking health information online: Does Wikipedia matter? *Journal of the American Medical Informatics Association*, 16(4):471–479, 2009.

[87] N. Limsopatham, C. Macdonald, and I. Ounis. Inferring conceptual relationships to improve medical records search. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, pages 1–8, 2013.

[88] N. Limsopatham, C. Macdonald, and I. Ounis. Learning to handle negated language in medical records search. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, pages 1431–1440. ACM, 2013.

[89] J. Lin and C. Dyer. *Data-Intensive Text Processing with MapReduce*. Morgan & Claypool, San Rafael, CA, 2010.

[90] D. A. Lindberg and B. L. Humphreys. 2015-The future of medical libraries. *New England Journal of Medicine*, 352(11):1067–1070, 2005.

[91] I. Mani. *Automatic Summarization*, Volume 3. John Benjamins Publishing, 2001.

[92] K. McKibbon and D. B. Fridsma. Effectiveness of clinician-selected electronic information resources for answering primary care physicians? Information needs. *Journal of the American Medical Informatics Association*, 13(6):653–659, 2006.

[93] K. McKibbon, R. B. Haynes, C. J. Walker Dilks, M. F. Ramsden, N. C. Ryan, L. Baker, T. Flemming, and D. Fitzgerald. How good are clinical MEDLINE searches? A comparative study of clinical end-user and librarian searches. *Computers and Biomedical Research*, 23(6):583–593, 1990.

[94] J. Metzger and J. Rhoads. Summary of Key Provisions in Final Rule for Stage 2 HITECH Meaningful Use. Falls Church, VA. *Computer Sciences Corp*, 2012. http://assets1.csc.com/health_services/downloads/CSC_Key_Provisions_of_Final_Rule_for_Stage_2.pdf

[95] M. Mitra, A. Singhal, and C. Buckley. Improving automatic query expansion. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 206–214. ACM, 1998.

[96] B. T. Mynatt, L. M. Leventhal, K. Instone, J. Farhat, and D. S. Rohlman. Hypertext or book: Which is better for answering questions? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 19–25. ACM, 1992.

[97] P. Nakov, A. S. Schwartz, E. Stoica, and M. A. Hearst. Biotext team experiments for the TREC 2004 Genomics Track. http://trec.nist.gov/pubs/trec13/papers/ucal-berkeley.geo.pdf

[98] C. Neylon. Science publishing: Open access must enable open use. *Nature*, 492(7429):348–349, 2012.

[99] D. T. Nicholson. *An Evaluation of the Quality of Consumer Health Information on Wikipedia*. PhD thesis, Oregon Health & Science University, 2006.

[100] P. Pluye and R. Grad. How information retrieval technology may impact on physician practice: An organizational case study in family medicine. *Journal of Evaluation in Clinical Practice*, 10(3):413–430, 2004.

[101] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 275–281. ACM, 1998.

[102] K. Purcell, J. Brenner, and L. Rainie. Search engine use 2012. 2012. http://www.pewinternet.org/Reports/2012/Search-Engine-Use-2012.aspx

[103] Y. Qi and P.-F. Laquerre. Retrieving medical records with sennamed: NEC Labs America at TREC 2012 Medical Records Track. In *The Twenty-First Text REtrieval Conference Proceedings (TREC 2012) [NIST Special Publication: SP 500-298]*. National Institute of Standards and Technology-NIST, 2012. http://trec.nist.gov/pubs/trec21/papers/sennamed.medical.final.pdf

[104] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford Okapi at TREC-3. In *Overview of the Third Text REtrieval Conference (TREC-3)*. National Institute of Standards and Technology, 1994.

[105] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 232–241. Springer-Verlag New York, 1994.

[106] C. Safran, M. Bloomrosen, W. E. Hammond, S. Labkoff, S. Markel-Fox, P. C. Tang, and D. E. Detmer. Toward a national framework for the secondary use of health data: An American Medical Informatics Association white paper. *Journal of the American Medical Informatics Association*, 14(1):1–9, 2007.

[107] G. Salton. Developments in automatic text retrieval. *Science*, 253(5023):974–980, 1991.

[108] G. Salton and C. Buckley. Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41(4):288–297, 1990.

[109] G. Salton, E. A. Fox, and H. Wu. Extended Boolean information retrieval. *Communications of the ACM*, 26(11):1022–1036, 1983.

[110] G. Salton and M. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, 1983.

[111] E. W. Sayers, T. Barrett, D. A. Benson, E. Bolton, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, S. Federhen, et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 39(suppl 1):D38–D51, 2011.

[112] K. Seki, J. C. Costello, V. R. Singan, and J. Mostafa. TREC 2004 Genomics Track experiments at IUB. In *TREC*, 2004. http://trec.nist.gov/pubs/trec13/papers/indianau-seki.geo.pdf

[113] H. C. Sox. Medical journal editing: Who shall pay? *Annals of Internal Medicine*, 151(1):68–69, 2009.

[114] M. H. Stanfill, M. Williams, S. H. Fenton, R. A. Jenders, and W. R. Hersh. A systematic literature review of automated clinical coding and classification systems. *Journal of the American Medical Informatics Association*, 17(6):646–651, 2010.

[115] D. R. Swanson. Historical note: Information retrieval and the future of an illusion. *Journal of the American Society for Information Science*, 39(2):92–98, 1988.

[116] A. Taylor. A study of the information search behaviour of the millennial generation. *Information Research: An International Electronic Journal*, 17(1):n1, 2012. http://informationr.net/ir/17-1/paper508.html

[117] R. H. Thiele, N. C. Poiro, D. C. Scalzo, and E. C. Nemergut. Speed, accuracy, and confidence in Google, Ovid, PubMed, and UpToDate: Results of a randomised trial. *Postgraduate Medical Journal*, 86(1018):459–465, 2010.

[118] A. Turpin and W. Hersh. Do clarity scores for queries correlate with user performance? In *Proceedings of the 15th Australasian Database Conference-Volume 27*, pages 85–91. Australian Computer Society, 2004.

[119] H. Turtle and W. B. Croft. Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems (TOIS)*, 9(3):187–222, 1991.

[120] A. J. van Deursen. Internet skill-related problems in accessing online health information. *International Journal of Medical Informatics*, 81(1):61–72, 2012.

[121] R. Van Noorden. The true cost of science publishing. *Nature*, 495(7442):426–429, 2013.

[122] D. Hiemstra and W. Kraaij. A language-modeling approach to trec. In E. Voorhees and D. Harman, editors, *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press, Cambridge, MA, 2005.

[123] E. M. Voorhees. The TREC Medical Records Track. In *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics*, pages 239–246. ACM, 2013.

[124] N. Walczuch, N. Fuhr, M. Pollmann, and B. Sievers. Routing and ad-hoc retrieval with the TREC-3 collection in a distributed loosely federated environment. *TREC-3*, 135–144.

[125] S. L. Weibel and T. Koch. The Dublin Core Metadata Initiative. *D-lib Magazine*, 6(12), 2000.

[126] J. I. Westbrook, E. W. Coiera, and A. S. Gosling. Do online information retrieval systems help experienced clinicians answer clinical questions? *Journal of the American Medical Informatics Association*, 12(3):315–321, 2005.

[127] W. J. Wilbur and Y. Yang. An analysis of statistical term strength and its use in the indexing and retrieval of molecular biology texts. *Computers in Biology and Medicine*, 26(3):209–222, 1996.

[128] B. M. Wildemuth, R. de Bliek, C. P. Friedman, and D. D. File. Medical students' personal knowledge, searching proficiency, and database use in problem solving. *Journal of the American society for Information Science*, 46(8):590–607, 1995.

[129] A. J. Wolpert. For the sake of inquiry and knowledge? The inevitability of open access. *New England Journal of Medicine*, 368(9):785–787, 2013.

[130] D. A. Zarin and T. Tse. Trust but verify: Trial registration and determining fidelity to the protocol. *Annals of Internal Medicine*, 159(1):65–67, 2013.

[131] D. A. Zarin, T. Tse, R. J. Williams, R. M. Califf, and N. C. Ide. The clinicaltrials.gov results database? Update and key issues. *New England Journal of Medicine*, 364(9):852–860, 2011.

[132] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems (TOIS)*, 22(2):179–214, 2004.

[133] Z. Zheng, S. Brady, A. Garg, and H. Shatkay. Applying probabilistic thematic clustering for classification in the TREC 2005 Genomics Track. In *TREC*, 2005.

[134] X. Zhou, X. Hu, and X. Zhang. Topic signature language models for ad hoc retrieval. *IEEE Transactions on Knowledge and Data Engineering,*, 19(9):1276–1287, 2007.

[135] D. Zhu and B. Carterette. An adaptive evidence weighting method for medical record search. In *Proceedings of the 36th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1025–1028. ACM, 2013.